Gabor filter tutorial pdf download full form pdf

This information was used to infer the location where an individual may be found. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 489–496. 110, 30–47. They applied fuzzy integral techniques to combine the outputs of two different SVM classifiers, increasing the computational burden of the method. Most of the existing datasets contain very few classes (15 on average). (2014a) proposed a scale and shape invariant method for localizing complex spatiotemporal events in video sequences. Google Scholar Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. Moutzouris et al. Most of the work in human activity recognition assumes a figure-centric scene of uncluttered background, where the actor is free to perform an activity. (1994). Instead of explicitly using separate object detectors, they divided the frames into regions and treated each region as an object candidate. Generating such information is not an easy task. Google Scholar Kovashka, A., and Grauman, K. 5.3. Methods Based on Social Networking Social interactions are an important part of daily life. Even though their method performs well in recognizing human interaction, the lack of an intrinsic audio-visual relationship estimation limits the recognition problem. C., and Savarese, S. European Conference on Computer Vision (Zurich), 689–704. The new feature could act as complementary material to the low-level feature. (2004), was presented by Yan and Luo (2012). Fathi et al. Thus, HMMs are not suitable for recognizing more complex events, but rather an event is decomposed into simpler activities, which are easier to recognize. Usually, the detection of objects in a scene may help to better understand human activities as it may provide useful information about the ongoing event (Gupta and Davis, 2007). Z., and Torre, F. Figure 10. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 2847–2854. Representative stochastic approaches for action recognition. Unimodal approaches are appropriate for recognizing human activities based on motion features. In Aggarwal and Cai (1999), a new taxonomy was presented focusing on human motion analysis, tracking from single view and multiview cameras, and recognition of human activities. J., and Levine, M. "Action recognition with motion-appearance vocabulary forest," in Proc. "A biologically inspired system for action recognition," in Proc. Also, their method used different sets of hidden states for audio and visual information. (2009) presented a review of state-of-the-art affective recognition methods that use visual and audio cues for recognizing spontaneous affective states and provided a list of related datasets for human affective expression recognition. A., and Pour, P. Training and validation methods still suffer from limitations, such as slow learning rate, which gets even worse for large scale training data, and low recognition rate. Google Scholar Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2014) reduced the required amount of pose data using a fixed length vector of more informative motion features (e.g., location and velocity) for each skeletal point. In the simple case, a human activity recognition system may recognize the underlying activity by taking into account only the visual information. The work of Shotton et al. Vrigkas et al. Google Scholar Hoai, M., and Zisserman, A. "Modeling mutual context of object and human pose in human-object interaction activities," in Proc. Google Scholar Ryoo, M. S. Vol. 70, 70–91. Rohrbach, R., Todorovic, S., and Zhu, S. (2012) represented the concept of social interactions as an oriented graph using an influence model to identify human interactions. C., Wang, H., and Fei-Fei, L. Google Scholar Martinez, H. To overcome this drawback, several approaches employing transfer learning (Lampert et al., 2009; Kulkarni et al., 2014), multitask learning (Evgeniou and Pontil, 2004; Salakhutdinov et al., 2011), and semantic/discriminative attribute learning (Farhadi et al., 2009; Jayaraman and Grauman, 2014) were proposed to automatically generate and handle the most informative attributes for human activity classification. Each link measured the influence of a person over another. Attributes describe specific properties of human actions, while parts of actions, which were obtained from objects and human poses, were used as bases for learning complex activities. "Recognizing human emotions from body movement and gesture dynamics," in Proc. Google Scholar Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. P., Dollár, P., Lin, D., Anderson, D. Many approaches, which employ background subtraction (Sigal et al., 2012a) or assume fixed limb lengths and uniformly distributed rotations of body parts (Burenius et al., 2013), have been proposed to reduce the complexity of the 3D space. Such social attributes may play an important role in analyzing social behaviors, which are the key to social engagement. "Mining actionlet ensemble for action recognition with depth cameras," in Proc. Then, each of these two categories is further analyzed into sub-categories depending on how they model human activities. "Real-time human pose recognition in parts from single depth images," in Proc. It can be seen as an interpretation of human speech, facial expressions, gestures, and movements. The main idea is to find the neighboring features around the detected interest points, quantize them, and form a vocabulary. Hussain et al. The articulated human body is usually represented as a tree-like structure, thus locating the global position and tracking each limb separately is intrinsically difficult, since it requires exploration of a large state space of all possible translations and rotations of the human body parts in 3D space. "A survey on human motion analysis from depth data," in Time-of-Flight and Depth Imaging, Lecture Notes in Computer Science, Vol. Google Scholar Tran, D., Yuan, J., and Forsyth, D. Google Scholar Efros, A. Figure 2. Google Scholar Livne, M., Sigal, L., Troje, N. doi:10.1109/PROC.2002.801448 CrossRef Full Text | Google Scholar Escalera, S., Baró, X., Vitrià, J., Radeva, P., and Raducanu, B. A unified tree-based framework for joint action localization, recognition and segmentation. Xiong et al. Most of the existing approaches for modeling group activities and social interactions between different persons usually exploit contextual information from the scenes. Motivated by this fact, human activity recognition systems aim to correctly classify input data into its underlying activity category. The new representation of fused features was used to recognize complex social events. Cornell University Library. A survey on still image based human action recognition. Comparison of unimodal methods. A combination of activity recognition and localization was presented by Chen and Grauman (2012). IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1290–1297. To assess their method, the authors introduced a large dataset with >60,000 static images obtained from the web, called web image dataset for event recognition (WIDER). An event can be described by different types of features that provide more and useful information. (2010) proposed a human activity recognition system by taking advantage of the auditory information of the video sequences of the HOHA dataset (Laptev et al., 2008) and used late fusion techniques for combining audio and visual cues. (2013a). X., Xiong, C., and Zhu, S. Chakraborty et al. Recent developments in human motion analysis. In: et al. 117, 633–659. The number and type of different modalities that can be used for analyzing human activities is an important question. 15, 113–129. doi:10.1007/s11263-013-0662-8 CrossRef Full Text | Google Scholar Wang, H., Ullah, M. M., and Schmid, C. Google Scholar Sargin, M. Google Scholar Lin, Z., Jiang, Z., and Davis, L. S., and Gall, J. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3273–3280. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Miami Beach, FL), 2929–2936. (2015) recognized complex video events and group activities from aerial shoots captured from unmanned aerial vehicles (UAVs). (2014) used a 3D space-time volume representation of human actions obtained from super-voxels to understand sport activities. Most of the existing studies in this field failed to efficiently describe human activities in a concise and informative way as they introduce limitations concerning computational issues. Google Scholar Lampert, C. C., Oreifej, O., and Masci, J., Gambardella, L. A recent review on 3D pose estimation and activity recognition was proposed by Holte et al. Conference on Human Motion: Understanding, Modeling, Capture and Animation (Rio de Janeiro), 271–284. These features were clustered using K-means to build a hierarchical template tree representation of each action. Moreover, the authors were able to estimate unseen poses using a hierarchical manifold search method. Therefore, automatic affective recognition systems should reduce the effort for selecting the proper affective label to better assess human emotions. The proposed scheme was based on random forests, which could select samples of spatiotemporal volumes in a video that characterize an action. ACM International Conference on Knowledge Discovery and Data Mining (Seattle, WA), 109–117. Google Scholar Ko, T., and Nevatia, R. To model this descriptor, a Bag-of-Words (BoW) technique is employed, whereas, classification of activities is performed using relevant vector machines (RVM) (Tipping, 2001). Human activity understanding has become one of the most active research topics in computer vision. A comprehensive review of existing human activity classification benchmarks was also presented and we examined the challenges of data acquisition to the problem of understanding human activity. However, these datasets still remain popular for human activity classification, as they provide a good evaluation criterion for many new methods. doi:10.1109/TMM.2013.2267205 CrossRef Full Text | Google Scholar Abouelenien, M., Perez-Rosas, V., Mihalcea, R., and Burzo, M. (2011) modeled 3D human poses and performed human activity recognition from depth images by mapping the pose estimation problem into a simpler pixel-wise classification problem. doi:10.1007/s00138-012-0450-4 CrossRef Full Text | Google Scholar Robertson, N., and Reid, I. Affective Comput. Dempster-Shafer theory (Shafer, 1976) was employed for fusing the different modalities, while SVM was used for classification. Asian Conference on Computer Vision (Singapore), 428–443. doi:10.1016/j.cviu.2012.01.003 CrossRef Full Text | Google Scholar Li, Z., Yu, X., R., and Corso, J. The recognition processes could be applied in real time using the incremental covariance update and the on-demand nearest neighbor classification schemes. (2011) focused on automatically tracking and recognizing players' positions (i.e., attacker and defender) in sports' videos. 107, 219–238. "Talking heads: detecting humans and recognizing their interactions," in Proc. A., and Shah, S. 7, 530–542. doi:10.1016/j.patcog.2013.10.019 CrossRef Full Text | Google Scholar Yan, X., and Luo, Y. Rapkis et al. Escalera et al. 108 (Special Issue on Vision for Human-Computer Interaction), 116–134. doi:10.1109/TIFS.2014.2344448 CrossRef Full Text | Google Scholar IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 4777–4785. (2015) represented skeletal joints as points on the product space. Google Scholar Venulapalli, R., Arrate, F., and Chellappa, R. "Action snippets: how many frames does human action recognition require?," in Proc. Google Scholar Kong, Y., and Fu, Y. Background subtraction (Elgammal et al., 2002) and human tracking (Liu et al., 2010) are usually used as a part of this process; (ii) human activity modeling (e.g., feature extraction; Laptev, 2005) is the step of extracting the necessary information that will help in the recognition step; and (iii) human activity classification is the step where a probe video sequence is classified in one of the classes of the activities that have been defined before building the system. (2003) proposed a hierarchical action categorization hierarchy. 28, 976–990. Soleymani et al. How to determine whether a human activity classification system provides the best performance? 4.3. Rule-Based Methods Rule-based approaches determine ongoing events by modeling an activity using rules or sets of attributes that describe an event. Depending on their complexity, human activities are categorized into: (i) gestures; (ii) atomic actions; (iii) human-to-object or human-to-human interactions; (iv) group actions; (v) behaviors; and (vi) events. The proposed model, also known as infinite hidden conditional random field (iHCRF), was employed to recognize emotional states, such as pain and agreement, and disagreement from non-verbal multimodal cues. Based on the technique of BoW, each action was presented by a histogram of visual words, whereas their approach was robust to background clutter. 48, 70–80. Finally, a thorough analysis of the ontologies for human behavior recognition from the perspective of data and knowledge representation was presented by Rodríguez et al. "ACTION MACH: a spatio-temporal maximum average correlation height filter for action recognition," in Proc. Kviatkovsky et al. 25, 71–84. Google Scholar Wu, X., Xu, D., Duan, L., and Luo, J. Conference on Human Factors in Computing Systems (Austin, TX), 1737–1746. Google Scholar Jiang, Z., Lin, Z., and Davis, L. Thus, a simultaneous inference of actions and actors is required. J., Darrell, T., et al. "Deeply learned attributes for crowded scene understanding," in Proc. In what level can the system reach the human ability of recognizing a human activity? An action descriptor of histograms of interest points, relying on the work of Schuldt et al. "Discovering discriminative action parts from mid-level video representations," in Proc. They analyzed the properties of directly used affect annotations in classification models, and proposed a method for transforming such annotations to build more accurate models. Facial Action Coding System (FACS): Manual. Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3337–3344. (1981). To achieve such a high computational speed, the authors used random walk sub-sampling methods. Context-sensitive learning for enhanced audiovisual emotion classification. James and Siebe (2007) proposed a survey for multimodal human computer interaction focusing on affective interaction methods from poses, facial expressions, and speech. doi:10.1109/T-AFFC.2011.40 CrossRef Full Text | Google Scholar Mikolajczyk, K., and Uemura, H. A partial least squares approach was used for learning the representation of action features, which is then fed into an SVM classifier. A hierarchical approach was followed by Jhuang et al. "Objects in action: an approach for combining action understanding and object perception," in Proc. Baxter et al. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. The type of interaction was recognized by assigning social roles to each person. Google Scholar Hoai, M., Lan, Z. Y., Lee, S., Heo, Y. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 1100–1109. They consider an activity as the 3D space-time volume, consisting of concatenation of 2D spaces in time. Kong and Fu (2014) addressed the problem of human interaction classification from subjects that lie close to each other. "Atomic action features: a new feature for action recognition," in Proc. Canonical correlation analysis (CCA) (Hardoon et al., 2004) was widely studied in the literature as an effective way for fusing data at the feature level (Sun et al., 2005; Wang et al., 2011c; Rudovic et al., 2013). European Conference on Computer Vision (Firenze), 291–300. The estimation of human pose is also very sensitive to several factors, such as illumination changes, variations in view-point, occlusions, background clutter, and human clothing. Google Scholar Metallinou, A., Lee, A. S., and Narayanan, S. S., and Fernández-Caballero, A. (2015) utilized CNNs to hierarchically combine information from different visual channels. A survey of advances in vision-based human motion capture and analysis. In particular, human actions are represented by vectors of dissimilarities and a set of prototype actions is built. Also, the construction of a visual model for learning and analyzing human movements in real time with inadequate benchmark datasets for evaluation is challenging tasks. "Can humans fly? One important aspect of human behavior recognition is the choice of proper features, which can be used to recognize behavior in applications, such as gaming and physiology. "Spatiotemporal deformable part models for action detection," in Proc. Deep learning has also been used by Gan et al. Gaussian processes were used as an online probabilistic regressor for this task using sparse representation of data for reducing computational complexity. The problem of temporal segmentation and event recognition was examined by Hoai et al. Google Scholar Liu, J., Luo, J., and Shah, M. (2012b) leveraged the shared structural analysis for activity recognition. M., Nickisch, H., and Harmeling, S. A vocabulary was learned from these features and SVM was used for classification. Much focus has also been given in recognizing human activities from real life videos, such as movies and TV shows, by exploiting scene contexts to localize activities and understand human interactions (Marszalek et al., 2009; Patron-Perez et al., 2012; Bojanowski et al., 2013; Hoai and Zisserman, 2014). (2003). Learning how to recognize complex human activities has been one of the tough obstacles in this field. Complexity reduction is achieved by associating intermediate features and class labels, is a necessary aspect for bridging the semantic gap. "Actions in context," in Proc. Google Scholar Jhuang, B., Martinez, R., Valstar, M. I., and Hung, T. Graphical models have been widely used in modeling 3D human poses. J. Yan and Luo (2012) also proposed a novel action descriptor based on spatial temporal interest points (STIP) (Laptev, 2005). In Section 6, we provide a categorization of human activity classification datasets and discuss some future research directions. 34, 1995–2006. Most of the existing approaches represent human activities as a set of visual features extracted from video sequences or still images and recognize the underlying activity label using several classification models (Kong et al., 2014a; Wang et al., 2014). Often, the observation sequence may be ignored due to normalization leading to the label bias problem (Lafferty et al., 2001). "Multimodal deep learning," in Proc. (2006), Chaudhry et al. The gap of a complete representation of human activities and the corresponding data collection and annotation is still a challenging and unbridged problem. Kuehne et al. "Activity recognition using dynamic subspace angles," in Proc. E. 16, 2639–2664. (2013a) exploited different types of features, such as static and motion features, for recognizing unlabeled events from heterogeneous web data (e.g., YouTube and Google/Bing image search engines). Image Understanding 116, 648–660. "Action recognition via local descriptors and holistic features," in Proc. Y. doi:10.1109/TMM.2013.2293060 CrossRef Full Text | Google Scholar Wang, Y., and Mori, G. Similarly, Chen et al. "Vector boosting for rotation invariant multi-view face detection," in Proc. IEEE International Conference on Computer Vision, pages (Sydney, NSW), 2696–2703. Google Scholar Poppe, R. "Detecting activities of daily living in first-person camera views," in Proc. Their method was able to distinguish between easy and difficult examples for representing the different classes and recognizing the underlying activity is difficult. Although the understanding of human activities may benefit from affective state recognition, the classification process is extremely challenging due to the semantic gap between the low-level features extracted from video frames and high-level concepts, such as emotions, that need to be identified. doi:10.1016/j.patcog.2011.12.028 CrossRef Full Text | Google Scholar Turaga, P. A hierarchical structure, which is called the short-product network, was used by Amer and Todorovic (2012). A common problem in estimating human pose is the high-dimensional space (i.e., each limb may have a large number of degrees of freedom that need to be estimated simultaneously). We classify unimodal methods into four broad categories: (i) space-time, (ii) stochastic, (iii) rule-based, and (iv) shape-based approaches. Schuldt et al. (2011). The main problem is how we can ensure the continuity of the motion along time as an action occurs uniformly or non-uniformly within a video sequence. (2012a) for human action recognition with fuzzy distances, to achieve time invariant action representation. Although there exists a plethora of benchmark activity recognition datasets in the literature, we have focused on the most widely used ones with respect to the database size, resolution, and usability. doi:10.1023/B:STCO.0000035301.49549.88 CrossRef Full Text | Google Scholar Snoek, C. Machine learning techniques that incorporate knowledge-driven approaches may be vital for human activity modeling and recognition in unconstrained environments, where data may not be adequate or may suffer from occlusions and changes in illuminations and view point. Google Scholar Nicolaou, M. Google Scholar Maji, S., Bourdev, L. "Human action recognition by representing 3D skeletons as points in a lie group," in Proc. They also proposed a new motion descriptor called "divergence-curl-shear descriptor," which is able to capture the hidden properties of flow patterns in video sequences. "Space-time tree ensemble for action recognition," in Proc. IEEE. Special Issue on Multimodal Human-Computer Interaction, Invited Paper, Vol. Lett. (2013) explored the effect of the affective feature variations over time on the classification of affective states. However, their method cannot deal with data that contain complex backgrounds, and due to the down-sampling of the original data the audio-visual synchronization may be lost. Kulkarni et al. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 1110–1118. M., Sastry, S. doi:10.1016/j.imavis.2012.07.003 CrossRef Full Text | Google Scholar Bousmalis, K., Zafeiriou, S., Morency, L. M., and Schmidhuber, J. Because it provides information about the identity of a person, their personality, and psychological state, it is difficult to extract. IEEE 92, 495–513. The final classification was performed by a multiclass SVM classifier. Google Scholar Ryoo, M. S., and Mori, G. 24, 971–981. Motion features were extracted from a video sequence by Messing et al. The non-complex backgrounds and the non-intraclass variations in human movements make these datasets non-applicable for real world applications. The geometry between different body parts was taken into account, and a 3D representation of human skeleton was proposed. The perception is performed into the dissimilarity space using sparse representation-based classification. Modeling local behavior for predicting social interactions towards human tracking. However, several intraclass variations caused by missing data, partial occlusion, and the sort duration of actions in time may harm the recognition accuracy. (2011a) focused on tracking dense sample points from video sequences using optical flow based on HCRFs for object recognition. doi:10.1155/S1110865703211010 CrossRef Full Text | Google Scholar Wu, C., Zhang, J., Savarese, S., and Saxena, A. Sun. (2013a) proposed a combination of shape and appearance descriptors to represent local features for human pose estimation. "Bimodal log-linear regression for fusion of audio and visual features," in Proc. Google Scholar Lan, T., Wang, Y., Yang, W., Robinovitch, S. Inform. Our goal is not only to present a new classification for the human activity recognition methods but also to compare different state-of-the-art studies and understand the advantages and disadvantages of each method. Example of interacting persons. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 46–55. Shape features were represented as high-dimensional non-linear structures such as hierarchical trees and kernels. A new type of feature called the "hankelet" was presented by Li et al. The proposed multimodal feature fusion techniques do not incorporate the special characteristics of each modality and the level of abstraction for fusing. Neurocomputing 87, 51–61. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (Los Alamitos, CA), 58–65. Google Scholar Andriluka, M., and Sigal, L. "Active learning of an action detector from untrimmed videos," in Proc. 24, 170–177. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. Activity recognition using the velocity histories of tracked keypoints," in Proc. They used multiple-instance formulation in conjunction with an MRF model and were able to represent human activities with a bag of Markov chains obtained from STIP and salient region feature selection. 34, 1549–1562. Google Scholar Pantic, M., Pentland, A., Nijholt, A., and Huang, T. International Conference on Affective Computing and Intelligent Interaction and Workshops (Amsterdam: De Rode Hoed), 1–6. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 588–595. A survey on ontologies for human behavior recognition. In addition, the system should work regardless of any external factors. (2012) employed several hierarchical classifiers models from neural networks to HMMs and their combinations to recognize audio-visual emotional levels of valence and arousal rather than emotional labels, such as anger and kindness. F., Delaitre, V., Gupta, A., Efros, A., Shao et al. V., and Schiele, B. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3177–3184. A general method for human activity recognition in video. Multimedia 16, 1525–1535. 6974 (Memphis, TN), 195–204. "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition," in Proc. Loose-limbed people: estimating 3D human pose and motion using non-parametric belief propagation. (2010) combined actions and human poses together, treating poses as latent variables, to infer the action label in still images. A multiview activity recognition method was presented by Li and Zickler (2012), where descriptors from different views were connected together to construct a new augmented feature that contains the transition between the different views. doi:10.1109/JSTSP.2012.2193556 CrossRef Full Text | Google Scholar Holte, M. Google Scholar Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. Decomposition of human activities. Theodorakopoulos et al. Google Scholar Kviatkovsky, I., Rivlin, E., and Shimshoni, I. (2001). International Conference on Digital Image Computing: Techniques and Applications (Sydney, NSW), 288–293. 31, Clausi, D., and Zelek, J. doi:10.1007/s00138-013-0521-1 CrossRef Full Text | Google Scholar Marszalek, M., Laptev, I., and Schmid, C. "Combining skeletal pose with local motion for human activity recognition," in Proc. K., and Calvo, R. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 886–893. Google Scholar Zeiler, D. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Miami Beach, FL), 1778–1785. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Miami Beach, FL), 951–958. However, this approach is incapable of discriminating between different hands. Google Scholar Kong, Y., Lee, H., and Provost, E. This type of feature, which was formed by short tracklets, along with a BoW approach, was able to recognize actions under different viewpoints without requiring any camera calibration. doi:10.1109/TCSVT.2008.2005594 CrossRef Full Text | Google Scholar Urtasun, R., and Darrell, T. doi:10.1109/TIFS.2011.2175921 CrossRef Full Text | Google Scholar Iosifidis, A., Tefas, A., and Pitas, I. Understanding both the actor and the action may be vital for real life applications, such as robot navigation and patient monitoring. doi:10.1162/089976604323181472 CrossRef Full Text | Google Scholar Vrigkas, M., Nikou, C., and Kakadiaris, I. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 3147–3155. Action recognition was performed by a supervised learning algorithm. (2012b). IEEE Computer Society Conference on Computer Vision and Pattern Recognition (San Francisco, CA), 17–24. A preprocessing step prior to the recognition process was adopted to address several limitations of frame capturing, such as low resolution, camera motion, and occlusions. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3249–3256. "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in Proc. Affective methods represent human activities according to emotional communications and the affective state of a person (Liu et al., 2011b; Martinez et al., 2014). Table 4 summarizes human activity recognition datasets, categorizing them into seven different categories. Google Scholar Salakhutdinov, R., Torralba, A., and Tenenbaum, J. Google Scholar Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L. "Joint motion segmentation and background estimation in dynamic scenes," in Proc. Y. Pattern Anal. 95, 15–48. "A database for fine grained activity detection of cooking activities," in Proc. Visualization of human actions with dense trajectories (row row). IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 2337–2344. However, this is challenging to obtain for real world situations, since affective events are expressed in a different manner by different persons or occur simultaneously with other activities and feelings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Miami Beach, FL), 951–958. Moreover, the problem of data association as parts of the human skeleton may vanish through the sequential layer propagation and back projection. An unsupervised method for learning human activities from short tracklets was proposed by Gaidon et al. Google Scholar Gaidon, A., Harchaoui, Z., and Schmid, C. Google Scholar Tipping, M. An approach for group activity classification was introduced by Choi et al. Poppe (2010) characterized human activity recognition methods into two main categories, describing them as "top-down" and "bottom-up." On the other hand, Aggarwal and Ryoo (2011) presented a tree-structured taxonomy, where the human activity recognition methods were categorized into two big sub-categories, the "single layer" approaches and the "hierarchical" approaches, each of which have several layers of categorization. However, building successful models for human pose estimation is not straightforward (Pishchulin et al., 2013). doi:10.1016/j.patcog.2015.02.019 CrossRef Full Text | Google Scholar Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., and Ilic, S. They took advantage of the Poisson equation solution to efficiently describe an action by using spectral clustering between sequences of features and applying nearest neighbor classification to characterize an action. Affective computing studies model the ability of a person to express, recognize, and control his/her affective states in terms of hardware and software. "Learning a discriminative hidden part model for human action recognition," in Proc. "Annotation and processing of continuous emotional attributes: challenges and opportunities," in Proc. Google Scholar Yu, S., Krim, H., and Norris, L. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. The authors were interested in classifying social activities of daily life, such as birthdays and weddings. Google Scholar Ikizler-Cinbis, N., and Sclaroff, S. Video event detection: from subvolume localization to spatiotemporal path search. IEEE International Conference on Advanced Video and Signal Based Surveillance (Boston, MA), 48–55. Bousmalis et al. Google Scholar Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. However, their method can cope with video sequences of frontal view only. A Mathematical Theory of Evidence. A., and Xia, D. One way to cope with this limitation is to segment video sequences into smaller clips that contain sub-actions, using a hierarchical approach (Pirsiavash and Ramanan, 2014). Although this approach seems to have the advantages of both early and late fusion techniques, it also has a large computational burden due to the different levels of information processing. Table 4. International Conference on Pattern Recognition (Cambridge), 32–36. Google Scholar Palatucci, M., Pomerleau, D., Hinton, G. (2013) used deep belief networks (DBN) (Hinton et al., 2006) in both supervised and unsupervised manners to learn the most informative audio-visual features and classify human emotions in dyadic interactions. Furthermore, several datasets are not generic, but rather cover a specific set of activities, such as sports and simple actions, which are usually performed by one actor. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Minneapolis, MN), 1–8. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3161–3167. "Multi-view latent variable discriminative models for action recognition," in Proc. (2014) have also addressed the problem of multiview pose estimation. 31, 1775–1789. Google Scholar Andriluka, M., Pishchulin, L., Gehler, P. "A Bayesian framework for 3D human motion tracking from monocular image," in IEEE International Conference on Acoustics, Speech and Signal Processing (Dallas, TX), 1398–1401. Several existing datasets have reached their expected life cycle (i.e., methods on Weizmann and KTH datasets achieved 100% recognition rate). 22, 4286–4300. (2014) proposed a semisupervised framework for recognizing human actions combining different visual features. European Conference on Computer Vision, Lecture Notes in Computer Science, Vol. "Multi-agent event recognition in structured scenarios," in Proc. doi:10.1109/TPAMI.2012.219 CrossRef Full Text | Google Scholar Huang, Z. Image Process. The second approach is based on the dependency between data frames, which encode the information about an action and the behavior of people in the neighborhood. Google Scholar Rahmani, H., and Mian, A. Google Scholar Jiang, Y. "Random treewalk toward instantaneous 3D human pose estimation," in Proc. "Human context: modeling human-human interactions for monocular 3D pose estimation," in Proc. They extracted features from space-time descriptors and compared them against features of the target video. An important activity dataset is required for the development of a human activity recognition system. (2013) used motion compensation techniques to recognize atomic actions. International Conference on Machine Learning (Bellevue, WA), 689–696. Behavioral approaches aim at recognizing behavioral attributes, non-verbal multimodal cues, such as gestures, facial expressions, and auditory cues. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Portland, OR), 3618–3625. Most of the studies of human activity recognition are associated with facial expression recognition and/or pose estimation techniques. Many algorithms convey a wealth of information about solving this problem. S., and Shi, J. By treating attributes as latent variables, the authors were able to annotate and classify video sequences of social activities. "Realistic human action recognition with audio context," in Proc. That is, actions within the same class may be expressed by different people with different body movements, and actions between different classes may be difficult to distinguish as they may be represented by similar information. "Watch-n-patch: unsupervised understanding of actions and relations," in Proc. A behavior recognition system may provide information about the personality and psychological state of a person, and its applications vary from video surveillance to human-computer interaction. Google Scholar Park, H. (2012a). Each group was assigned a learned weight according to its importance in motion representation. IEEE 98, 1692–1715. (2012a), where a new feature type called "local occupancy pattern" was also proposed. Nicolaou et al. "Label-embedding for attribute-based classification," in Proc. Proposed hierarchical categorization of human activity recognition methods. A major focus in action recognition from still images or videos has been made in the context of scene appearance (Thurau and Hlavac, 2008; Yang et al., 2010; Maji et al., 2011). Although CRFs outperform HMMs in many applications, including bioinformatics, activity, and speech recognition, the construction of more complex models for human activity recognition may have good generalization ability but is rather impractical for real time applications due to the large number of parameter estimations and the approximate inference. Understanding transit scenes: a survey on human behavior-recognition algorithms. Thus, some basic questions arise about a human activity classification system: 1. Google Scholar Jayaraman, D., and Grauman, K. Representing activities with layers of human poses is computationally expensive due to many degrees of freedom. A tutorial on support vector regression. "Action recognition using context and appearance distribution features," in Proc. (2008) proposed a probabilistic approach based on GMMs for recognizing human emotions in dyadic interactions. Due to their Markovian nature, they must enumerate all possible observation sequences while capturing the dependencies between each state and its corresponding observation only. V., and Hager, J. doi:10.1016/j.neucom.2007.12.044 CrossRef Full Text | Google Scholar Kläser, A., Marszałek, M., and Schmid, C. R., and Huang, T. S., Bauckhage, C., and Gall, J. Moreover, the creation of such a dataset should correspond to real world scenarios. Selective spatio-temporal interest points. Human interaction categorization by using audio-visual cues. Google Scholar Castellano, G., Villalba, S. P., Yannakakis, G. "Joint inference of groups, events and human roles in aerial videos," in Proc. "Joint segmentation and classification of human actions in video," in Proc. IEEE International Conference on Computer Vision (Barcelona), 2003–2010. Thus, such an approach is limited by the fact that only specific interaction patterns can be successfully modeled and is sensitive in modeling complex social events. Atomic actions are movements of a person describing a certain motion that may be part of more complex activities (Ni et al., 2015). Multimodal methods combine features collected from different sources (Wu et al., 2013) and are classified into three categories: (i) affective, (ii) behavioral, and (iii) social networking methods. Google Scholar Lichtenauer, J., Valstar, J. Another obstacle that researchers must overcome is the lack of adequate benchmark datasets to test and validate the reliability, effectiveness, and efficiency of a human behavior recognition system. (2015) proposed a novel method for reducing dimensionality of human poses called "hierarchical temporal Laplacian eigenmaps" (HTLE). The ability of a human activity classification system to imitate humans' skill in recognizing human actions in real time is a future challenge to be tackled. A probabilistic Markov random field (MRF) (Prince, 2012) framework was used to classify and localize the activities in a scene. Google Scholar Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. In general, effective feature extraction is highly application dependent. Google Scholar Wang, S., Yang, Y., Ma, Z., Li, X., Pang, C., and Hauptmann, A. IEEE Aerosp. IEEE Winter Conference on Applications of Computer Vision (Steamboat Springs, CO), 626–633. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Portland, OR), 2555–2562. Activity representation with motion hierarchies. Wu et al. At each level of abstraction, they learned a hierarchical model with latent variables to group similar semantic attributes of each layer. Hoai and Zisserman (2014) proposed a learning method based on the context and the properties of a scene for detecting upper body positions and understanding the interaction of the participants in TV shows. Attribute regularization based human action recognition. "Recognizing action at a distance," in Proc. (2003) recognized human actions from low-resolution sports' video sequences using the nearest neighbor classifier, where humans are represented by windows of height of 30 pixels. Such a representation may be erroneous to partial occlusions and feature-to-object mismatching. Google Scholar Candamo, J., Shreve, M., Goldgof, D. To achieve this, all stages of human activity modeling and analysis are to be performed automatically, namely: (a) human activity detection and localization, where the challenge is to detect and localize a human activity in the scene. Table 2. Wang et al. (2015) for detecting and recognizing complex events in video sequences. doi:10.1109/TCSVT.2008.2005594 CrossRef Full Text | Google Scholar Perronnin, F., and Dance, C. Multiview action recognition has also been studied by Rahmani and Mian (2015). doi:10.1109/TPAMI.2014.16 PubMed Abstract | CrossRef Full Text | Google Scholar Nie, B. (2012) proposed a mid-level approach extracting spatiotemporal features and constructing clusters of trajectories, which could be considered as candidates of an action. To this end, several kinematic and part-occlusion constraints for decomposing human poses into separate limbs have been explored to localize the human body (Cherian et al., 2014). However, this strategy is time-consuming and requires deep supervision learning schemes, which may cause a potential loss of inter-modality correlation. These methods assume that not only the human body itself, but the objects surrounding it, may provide evidence of the underlying activity. The intuition behind this approach is a psycholinguistics phenomenon, where randomizing letters in the middle of words has almost no effect on understanding the underlying word if and only if the first and the last letters of this word remain unchanged (Rawlinson, 2007). First, the authors applied human-to-object interaction techniques to identify the area of interest, then used this context-based information to train a conditional random field (CRF) model (Lafferty et al., 2001) and identify the underlying action. They considered the problem as two different tasks. Several challenges that correspond to the ability of a classification system to generalize under external factors, such as variations in human poses and different data acquisition, are still open issues. The problem of behavioral mimicry in social interactions was studied by Bilakhia et al. However, their method uses features that could be characterized by non-visual properties (Ryoo and Aggarwal, 2011). Ahead of Print. To better understand the behavioral mimicry, the authors suggested that... 2009–2016. 20, 1.1–1.19. Human action recognition using spatio-temporal features with sparse coding. (2013) proposed a probabilistic framework for multimodal affect recognition, which is also capable of handling incomplete information. Google Scholar Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Portland, OR), 2555–2562. The audio-visual synchronization... The classification of human activities, which are complex... Shabani et al. Google Scholar Wang, H., and Schmid, C. "Action recognition with improved trajectories," in Proc. The problem of human behavior understanding... International Conference on Computer Vision (Barcelona), 2003–2010. However, such an approach...

support vector machine (SVM) classifiers (Bishop, 2006). Despite the vast development of pose estimation algorithms, the problem still remains challenging for real time applications. Motivated by this fact, Gupta and Davis (2007) proposed a Bayesian approach that encodes object detection and localization for understanding human actions. eds M. This problem becomes more prominent as human emotions are continuous acts in time, and variations in human actions may be confusing or lead to subjective annotations. "A flow model for joint action recognition and identity maintenance," in Proc. Google Scholar Ma, S., Sigal, L., and Sclaroff, S. Google Scholar Ye, M., Zhang, Q., Wang, L., Zhu, J., Yangg, R., and Gall, J. "Early versus late fusion in semantic video analysis," in Proc. (2011) proposed a framework for fusing multimodal psychological features, such as heart and facial muscle activity, skin response, and respiration, for detecting and recognizing affective states. Google Scholar Niebles, J. "HMDB: a large video database for human motion recognition," in Proc. "Learning context for collective activity recognition," in Proc. On the other hand, the work of Livne et al. Google Scholar Holte, M. (2015) proposed a real-time emotion recognition system that modeled 3D facial expressions using random forests. "Recognizing human actions from still images with latent poses," in Proc. Google Scholar Rawlinson, G. K., and Rehg, J. doi:10.1109/34.868684 CrossRef Full Text | Google Scholar Ouyang, W., Chu, X., and Wang, X. 73, 82–98. "Actionness ranking with lattice conditional ordinal random fields," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Portland, OR), 3562–3569. Human activities, such as "walking" and "running," arise very naturally in daily life and are relatively easy to recognize. Space-time methods involve activity recognition methods, which represent human activities as a set of spatiotemporal features (Shabani et al., 2011; Li and Zickler, 2012) or trajectories (Li et al., 2012; Vrigkas et al., 2013). IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3169–3176. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 3156–3164. 6, 538–552. "Propagative Hough voting for human activity recognition," in Proc. Besides the vast amount of research in this field, a generalization of the learning framework is crucial toward modeling and understanding real world human activities. Toshev and Szegedy (2014) have also used deep learning for human pose estimation. Part-based motion descriptor image for human action recognition. M., Rosario, B., and Pentland, A. The correct annotation was given in each video under a semisupervised scheme. Sun and Nevatia (2013) treated video sequences as sets of short clips rather than a whole representation of actions. IEEE International Symposium on Multimedia (Berkeley, CA), 250–257. Often, human actions are highly correlated to the actor, who performs a specific action. To this end, they proposed joint hidden conditional random Fields (JHCRF) as a new classification scheme to take advantage of the multimodal data. Google Scholar Murray, R. Finally, social networking methods model the characteristics and the behavior of humans in several layers of human-to-human interactions in social events from gestures, body motion, and speech (Patron-Perez et al., 2012; Marín-Jiménez et al., 2014). (2011a) addressed the problem of recognizing actions by a set of descriptive and discriminative attributes. P., Collins, M., and Darrell, T. Google Scholar Lu, W. Such attributes may be descriptions of emotional states or cognitive states, such as activation, valence, and engagement. "Action recognition robust to background clutter by using stereo vision," in Proc. (2002). Acknowledgments This research was funded in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. Human actions were represented using a hierarchical AND/OR graph and dynamic programing was used to infer the class label. (2000). Low-cost devices, such as Microsoft Kinect and other RGB-D sensors, which provide 3D depth data of a scene, can efficiently leverage these limitations and produce a relatively good estimation of human pose, since they are robust to illumination changes and texture variations (Gao et al., 2015). Stat. 6974 (Memphis, TN), 568–577. Google Scholar Schindler, K., and Gool, L. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 1302–1311. Google Scholar Soomro, K., Zamir, A. (2015) addressed these limitations and proposed a general probabilistic framework for joint actor-action understanding while they presented a new dataset for actor-action recognition. The human activity categorization problem has remained a challenging task in computer vision for more than two decades. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 2467–2474. "Hierarchical recurrent neural network for skeleton based action recognition," in Proc. Real time classification and prediction of future actions was proposed by Morris and Trivedi (2011), where an activity vocabulary is learned through a three-step procedure. Preprocessing affective annotations may be detrimental for generating accurate and ambiguous affective models due to biased representations of affect annotation. (2014b) employed a fully connected CRF model to identify human behaviors, such as friendly, aggressive, and neutral. doi:10.1007/s11263-011-0493-4 CrossRef Full Text | Google Scholar Sigal, L., Isard, M., Haussecker, H., and Black, M. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3289–3296. "Action recognition by dense trajectories," in Proc. doi:10.1016/0146-664X(81)90073-3 CrossRef Full Text | Google Scholar Hardoon, D. Moreover, the field of psychology has attracted great interest in studying social interactions, as scientists may infer useful information about human behavior. Burgos-Artizzu et al. The work of Le et al. Google Scholar Thurau, C., and Hlavac, V. Google Scholar Bilakhia, S., Petridis, S., and Pantic, M. Maji et al. doi:10.1109/MAES.2007.327521 CrossRef Full Text | Google Scholar Reddy, K. (2014) proposed a structured temporal approach for daily living human activity recognition. Each video was represented by a set of action descriptors, which were put in correspondence. doi:10.1109/TPAMI.2012.67 PubMed Abstract | CrossRef Full Text | Google Scholar Yao, B., Jiang, X., Khosla, A., Lin, A. "Modeling the temporal extent of actions," in Proc. "3D pictorial structures for multiple view articulated pose estimation," in Proc. Action recognition of complex scenes with multiple subjects was proposed by Morariu and Davis (2011). IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 1600–1609. Google Scholar Tang, K. "Decision level fusion of domain specific regions for facial action recognition," in Proc. 6975 (Memphis, TN), 415–424. (2005). Metallinou et al. A probe video sequence was classified into its underlying activity according to its similarity with each representation in the codebook. A., and Ragheb, H. An algorithm that may recognize human actions in 3D space by a multicamera system was introduced by Holte et al. Google Scholar Li, B., Camps, O. (2007), where an input video was analyzed into several feature descriptors depending on their complexity. People watching: human actions as a cue for single view geometry. (2007) explored the dynamics of body movements to identify affective behaviors using time series of multimodal data. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1218–1225. Figure 8. This approach, called slow fusion, is a combination of the previous approaches and can be seen as a hierarchical fusion technique that slowly fuses data by successively passing information through early and late fusion levels. International Conference on Computer Vision Theory and Applications (Barcelona), 112–117. A vocabulary based approach has been proposed by Kovashka and Grauman (2010). B., Sapper, D. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 1836–1845. K., and Cai, Q. Google Scholar Sun, Q. "Finding actors and actions in movies," in Proc. (2012c) proposed a probabilistic model of two components. (2012) proposed a mid-level representation of human actions by computing local motion volumes in skeletal points extracted from video sequences and constructed a codebook of poses for identifying the action. Sigmoid and Gaussian envelopes were incorporated into the loss function of an SVM classifier, where the outliers are eliminated during the optimization process. Y., and Hauptmann, A. H., Xiong, C., and Corso, J. Human activity recognition from 3D data: a review. Shabani et al. They used a hierarchical clustering algorithm to represent videos with an unordered tree structure and compared all tree-clusters to identity the underlying activity. (2004) represented local events in a video using space-time features, while an SVM classifier was used to recognize an action. The extraction of low-level features that are focused on representing human motion is a very challenging task. The types of human poses, as well as the spatial relationship between the different human parts, were modeled. (2013) could efficiently extract and select the most informative multimodal features using deep learning to model emotional expressions and recognize the affective states of a person. This may lead to significant loss of the intermodality dependence, while it suffers from propagating the classification error across different levels of classifiers. The output of each layer, which corresponds to neighboring parts, is fused and fed as input to the next layer. Aggarwal and Xia (2014) recently presented a categorization of human activity recognition methods from 3D stereo and motion capture systems with the main focus on methods that exploit 3D depth data. Google Scholar Xu, C., Hsieh, S. IEEE International Conference on Computer Vision (Kyoto), 104–111. Affective Computing. Google Scholar Lan, T., Chen, T. "Human action recognition using distribution of oriented rectangular patches," in Proc. 1 (Washington, DC: IEEE Computer Society), 278–282. IEEE International Conference on Computer Vision (Barcelona), 113–120. (2007). "A discriminative model with multiple temporal scales for action prediction," in Proc. We discussed the different levels of representation of feature modalities and reported the limitations and advantages for each representation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (Providence, RI), 28–35. Fuzzy vector quantization and linear discriminant analysis were employed to recognize a human activity. A shared representation of human poses and visual information has also been explored (Ferrari et al., 2009; Singh and Nevatia, 2011; Yun et al., 2012). Commun. First, saliency maps were used for detecting and localizing events, and then deep learning was applied to the pretrained features for identifying the most important frames that correspond to the underlying event. Their approach relies on using deep neural networks (DNN) (Ciresan et al., 2012) for representing cascade body joint regressors in a holistic manner. Activity recognition and localization via a figure-centric model was presented by Lan et al. Although the proposed method could cope with missing information and variations in scene context, scale, and orientation of human poses, it is sensitive to localization of interacting members, which leads to erroneous predictions of the true class. Google Scholar Singh, V. In this survey, we differentiate between these two terms in the sense that the term "activity" is used to describe a sequence of actions that correspond to specific body motion. 4.4. Shape-Based Methods of modeling human activities. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3177–3184. "Do these interaction features" ... (Chakraborty, B., González, J., and Moeslund, T. It was based on the synergy of 3D space and time to construct a 4D descriptor of spatial temporal interest points and a local description of 3D motion features. The authors considered that a person could be detected by performing background subtraction techniques. Yang et al. (A) The discriminative hierarchical model for the recognition of human action from body poses. 5.2. Behavioral Methods Recognizing human behaviors from video sequences is a challenging task for the computer vision community (Candamo et al., 2010). (2014b) modeled crowded scenes as a graph of interacting persons. Proc. doi:10.1016/j.cviu.2013.01.013 CrossRef Full Text | Google Scholar Chaudhry, R., Ravichandran, A., Hager, G. The tracking problem was decomposed into smaller tasks by tracking all possible configurations of interaction effects, while the number of trackers was dynamically estimated. New York, NY: Cambridge University Press. When temporal grammars are used for action classification, the main problem consists in treating long video sequences due to the complexity of the models. (2014) used a bag of local spatiotemporal volume features approach to recognize and localize human actions from weakly labeled video sequences using multiple instance learning. Earlier approaches were based on describing actions by using dense trajectories. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 748–755. Background and foreground modeling using nonparametric kernel density for visual surveillance. Google Scholar Xiong, Y., Zhu, K., Lin, D., and Tang, X. (2011) recognized abnormal behaviors in human group activities. doi:10.1109/T-AFFC.2011.37 CrossRef Full Text | Google Scholar Song, Y., Morency, L. "Zero-shot learning with semantic output codes," in Proc. Karpathy et al. For the combination of the different modalities, the authors applied multitask deep learning. Google Scholar Jhuang, H., Gall, J., Zuffi, S., Schmid, C., and Black, M. An HMM was employed to encode human actions, whereas recognition was performed by searching for image features that represent an action. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 3642–3649. (2011) proposed a temporally asymmetric filtering for feature detection and activity recognition. IEEE International Workshop on Tracking Humans for the Evaluation of Their Motion in Image Sequences (THEMIS) (Kyoto), 1089–1096. S., and Udrea, O. "Actions as space-time shapes," in Proc. "A Bayesian computer vision system for modeling human interactions. International Conference on Document Analysis and Recognition, Vol. Jain et al. "Deeppose: human pose estimation via deep neural networks," in Proc. Forensics Secur. "Learning a non-linear knowledge transfer model for cross-view action recognition," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 5624–5632. A key issue in affective computing is accurately annotated data. "Youtube2text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in Proc. Pose-based human action recognition via sparse representation in dissimilarity space. S., Zeng, S. (2013b). Google Scholar Wang, L., Hu, W., and Tan, T. The authors were able to recognize activities such as a group of people talking or standing in a queue. Martinez et al. The significance of letter position in word recognition. J., and Schölkopf, B. doi:10.1109/TPAMI.2011.44 PubMed Abstract | CrossRef Full Text | Google Scholar Moutzouris, A., del Rincon, J. "Latent boosting for action recognition," in Proc. They introduced a new representation of human actions, called "hierarchical movemes," composed of different levels of coarse to fine-grained level granularity. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 2120–2127. A fundamental component of human behavior is the ability to interact with other people via their actions. Mixing body parts of different views may lead to ambiguities because of the multiple candidates of each camera view and false positive detections. Google Scholar Ramanathan, V., Liang, P., and Fei-Fei, L. Google Scholar Husain, M. By these means, they were able to capture the intraclass correlations between the learned attributes while they proposed a novel dataset of crowd scene understanding, called WWW crowd dataset. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 2211–2218. Selecting the proper features for human activity recognition has been proved a trial-and-error approach for many researchers in this area of research. Li et al. (2015) used the idea that one needs to relax the tight constraints of bounding box tracking, while they used a sliding window technique to track spatiotemporal paths maximizing the summation score. The approach of Fathi and Mori (2008) was based on mid-level motion features, which are also constructed directly from optical flow features. D., and Malik, J. (2012a) proposed a novel method for human behavior recognition based on multiview hidden conditional random fields (MV-HCRF) (Song et al., 2012b) and estimated the interaction of the different modalities by using kernel canonical correlation analysis (KCCA) (Hardoon et al., 2004). The quality of the input media that forms the dataset is one of the most important things one should take into account. Several feature descriptors, such as HOG3D (Kläser et al., 2008) and STIP (Laptev, 2005), are not able to sufficiently characterize human behaviors. This vector was given as input to an HMM, which was used for action classification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 3698–3706. Based on the first-order logic and probabilistic approaches, such as Markov networks, the authors were able to infer which event has occurred. (2015) also employed a hierarchical MRF model to represent segments of human actions by extracting super-voxels from different scales and automatically estimated the foreground motion using saliency features of neighboring super-voxels. Google Scholar Atrey, P. "Multimodal human behavior analysis: learning correlation and interaction across modalities," in Proc. All datasets are grouped by their associated category and by chronological order for each group. Conclusion In this survey, we carried out a comprehensive study of state-of-the-art methods of human activity recognition and proposed a hierarchical taxonomy for classifying these methods. Moreover, action categorization based on modeling the motion of parts of the human body was presented by Tran et al. Google Scholar Fouhey, D. 3D human poses have been taken into consideration in recent years and several algorithms for human activity recognition have been developed. Figure 6 summarizes their method using primitive rules for recognizing human actions. M., Ghodrati, A., and Tuytelaars, T. This property considers that the audio and visual features were a priori synchronized, while it increases the complexity of the model. Sensors 12, 1702–1719. This problem is generally known as zero-shot learning (Palatucci et al., 2009). "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1314–1321. doi:10.1109/TPAMI.2008.52 PubMed Abstract | CrossRef Full Text | Google Scholar Zhang, Z., Wang, C., Xiao, B., Zhou, W., and Liu, S. (1999). On the other hand, the term "behavior" is used to characterize both activities and events that are associated with gestures, emotional states, facial expressions, and auditory cues of a single person. Hidden conditional random fields. Proc. Google Scholar Moutzouris, A., Martinez-del-Rincon, J. Pose-based human action recognition via sparse representation in dissimilarity space. (2013) proposed an algorithm based on multilayer perceptrons, where each layer was fed with the output of the previous layer. Ji et al. Google Scholar Ke, Y., Sukthankar, R., and Hebert, M. Google Scholar Oliver, N. view invariant action classification. ACM Multimedia Conference (Barcelona), 789–792. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 612–619. Other optical flow-based methods which gained popularity were presented by Dalal et al. Classified Salient Features The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. 23, 412–424. Zeng et al. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3265–3272. D., McCallum, A., and Pereira, F. A multiview person identification was presented by Iosifidis et al. doi:10.1109/TPAMI.2010.214 PubMed Abstract | CrossRef Full Text | Google Scholar Westerveld, T., de Vries, A. Human body parts were handled as directional tree-structured representations and a regression tree was trained for each joint in the human skeleton. More specifically, Thurau and Hlavac (2008) represented actions by histograms of pose primitives, and n-gram expressions were used for action classification. 11, 206–224. Although much research has been focused on human activity recognition systems from video sequences, human activity recognition from static images remains an open and very challenging task. "Learning to share visual appearance for multiclass object detection," in Proc. Nevertheless, space-time features focus mainly on local spatiotemporal information. Google Scholar Gupta, A. Kembhavi, A., and Davis, L. (2011) discovered the action label in an unsupervised manner by learning features directly from video data. Machine recognition of human activities: a survey. Lan et al. Sparse Bayesian learning and the relevance vector machine. (2014) addressed the problem of identifying and localizing human actions using CRFs. The authors were able to distinguish between intentional actions and unknown motions that may happen in the surroundings by ordering video regions and detecting the actor of each action. doi:10.1109/TPAMI.2014.2328311 CrossRef Full Text | Google Scholar Yao, A., Gall, J., and Gool, L. 129, 15–26. Google Scholar Jain, M., Jegou, H., and Bouthemy, P. 8200. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. F., and Fleet, D. Section 3 presents the proposed categorization of human activities. Complex events were decomposed into simpler actions and modeled using a spatiotemporal CRF graph. N., and Hallam, J. The visual analysis of human movement: a survey. A., Pavlovic, V., and Pantic, M. 27, 1814–1825. Previous works on characterizing human behavior have shown great potential in this area. The authors modeled the behavior of humans in a scene using social roles in conjunction with modeling low-level actions and high-level events. The study of Yao and Fei-Fei (2012) modeled human poses for human-object interactions by introducing a mutual context model. Google Scholar Lu, W., Dellandrea, E., Tellez, B., and Chen, L. doi:10.1109/TIP.2013.2271850 CrossRef Full Text | Google Scholar Shabani, A. N., Gala, A., Kakadiaris, I. (2012) discussed the social behavior of mice. "Recognizing human actions by attributes," in Proc. (2015) presented a method for fast estimation of human pose with 1,000 frames per second. Pantic and Rothkrantz (2003) performed a complete study in human affective state recognition that incorporate non-verbal multimodal cues, such as facial and vocal expressions. "Recognizing human actions: a local SVM approach," in Proc. Audiovisual information fusion in human–computer interfaces and intelligent environments: a survey. A survey of human motion analysis using depth imagery. "Sum-product networks for modeling activities with stochastic structure," in Proc. A new method of feature fusion and its application in image recognition. "Action localization with tubelets from motion," in Proc. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (Geneva), 123–128. First, we categorize the human activity recognition methods into two main categories: (i) unimodal and (ii) multimodal activity recognition methods according to the nature of sensor data they employ. Neural Comput. Moreover, the association of human pose orientation with the poses extracted from different camera views is also used. Different action scenarios are given as input to the model, which is able to model the human interactions while being proposed by Oliver et al. In the sense of first-person scene understanding, Park and Shi (2015) were able to predict joint social interactions by modeling geometric relationships between groups of interacting persons. To this end, a study on how to produce highly informative affective labels has been proposed by Healey (2011). Google Scholar Cui, X., Liu, Q., Gao, M., and Metaxas, D. Google Scholar Zhou, W., and Zhang, Z. Each modality is separately analyzed and saliency scores are used for linear and non-linear fusing schemes. doi:10.1109/TPAMI.2011.228 CrossRef Full Text | Google Scholar Lan, T., Wang, Y., and Mori, G. Transp. Moreover, intra- and interclass similarities make the problem amply challenging. C., Busso, C., Garnicke, S. Figure 11. Hence, the recognition of complex activities may be crucial for understanding human behavior. Multimodal fusion for multimedia analysis: a survey. (2014) used a bag of visual-audio words scheme along with late fusion for recognizing human interactions in TV shows. Motion features were extracted forming a 3D HOG model. "Joint action recognition and pose estimation from video," in Proc. Google Scholar Alahi, A., Ramanathan, V., and Fei-Fei, L. IEEE International Conference on Computer Vision (Sydney, NSW), 1833–1840. International Joint Conference on Artificial Intelligence (Barcelona), 1237–1242. Jiang et al. Multimed. "Human action recognition by learning bases of action attributes and parts," in Proc. Google Scholar Quattoni, A., Wang, S., Morency, L. "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in Proc. A GMM was used for modeling human actions, and a transfer ranking technique was employed for recognizing unseen classes. (2011) proposed a recognition model based on SVMs for regression (SVR) (Smola and Schölkopf, 2004) for continuous prediction of multimodal emotional states, using facial expression, shoulder gesture, and audio cues in terms of arousal and valence (Figure 9). (2015) used a dynamic programming approach to recognize sequences of actions in untrimmed video sequences. Google Scholar Sadanand, S., and Corso, J. "Combining the right features for complex event recognition," in Proc. 103, 60–79. The work of Lu et al. (1997). Action recognition using depth cameras was introduced by Wang et al. A human action was modeled as a configuration of parts of image observations. M., and Watson, L. C., Mori, G., and Malik, J. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 1653–1660. G., Liu, Y., Heng, P. We also present the number of classes, actors, and video clips along with their frame resolution. A high-level representation of video sequences, called "action bank," was presented by Sadanand and Corso (2012). doi:10.1007/s11263-014-0710-z CrossRef Full Text | Google Scholar Fu, Y., Hospedales, T. The first step in developing a human activity recognition system is to acquire an adequate human activity dataset. doi:10.1016/j.cviu.2011.09.010 CrossRef Full Text | Google Scholar Chaquet, J. "A discriminative latent model of object classes and attributes," in Proc. The authors argued that visual information is not adequate for understanding human emotions, and thus additional information that describes the scene is needed. G., Ganes, H., and Pantic, M. A major family of methods relies on optical flow, which has proven to be an important cue. The individual strength of each modality may lead to better recognition results. Human interactions were addressed by Andriluka and Sigal (2012). Guo and Lai (2014) summarized all the methods for human activity recognition from still images and categorized them into two big categories according to the level of abstraction and the type of features each method uses. P., and Xue, Y. 18, 157–171. The proposed method was robust to subjects' poses and changes in the environment. (2012) proposed a propagative point-matching approach using random projection trees, which can handle unlabeled data in an unsupervised manner. IEEE International Conference on Computer Vision (Sydney, NSW), 3192–3199. The way that humans perform an activity depends on their habits, and this makes the problem of identifying the underlying activity quite difficult to determine. However, background subtraction can help overcome this limitation. Learning discriminative space-time action parts from weakly labelled videos. Finally, conclusions are drawn in Section 7. International Conference on Multimedia Retrieval (Toronto), 29–36. Google Scholar Chen, L., Duan, L., and Xu, D. doi:10.1109/TPAMI.2007.906843 CrossRef Full Text | Google Scholar Sultani, W., and Saleemi, I. (2013) addressed the problem of analyzing social interaction in human-object interactions and interactions. A typical example of a behavior recognition system is depicted in Figure 10. "Show and tell: a neural image caption generator," in Proc. Tables 2 and 3 provide a comprehensive comparison of unimodal and multimodal methods, respectively, and list the benefits and limitations of each family of methods. IEEE Winter Conference on Applications of Computer Vision (Steamboat Springs, CO), 220–226. (2013) used a feed-forward neural network using optical flow features. Each model captured a specific type of behavior. A. Chamela, A. Complex activities may be decomposed into other simpler activities, which are generally easier to recognize. C., Meng, U., and Schmidhuber, J. 56, 585–601. Forensics Secur 9, 1581–1591. (2014a). Human action recognition with multiple-instance Markov model. "Large-scale video classification with convolutional neural networks," in Proc. Unimodal human activity recognition methods identify human activities from data of one modality. Tran et al. The recognition process was performed over basketball game videos, where the players were first detected and tracked, generating a set of trajectories that are used to create a set of spatiotemporal events. Shape-based methods efficiently represent activities with high-level reasoning by modeling the motion of human body parts (Sigal et al., 2012b; Tran et al., 2012). "Affect detection and classification from the non-stationary physiological data," in Proc. Kong et al. IEEE International Conference on Computer Vision (Sydney, NSW), 913–920. Mach. Parts of the human body are described in 2D space as rectangular patches and as volumetric shapes in 3D space (see Figure 7). The most significant action paths were estimated by defining an action score. "Attribute learning for understanding unstructured social activity," in Proc. "Action recognition by hierarchical sequence summarization," in Proc. However, the ability to recognize the underlying class only from motion is on its own a challenging task. "Recognizing realistic actions from videos in the wild," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 2264–2273. IEEE International Conference on Computer Vision (Sydney, NSW), 1331–1338. A local 3-D motion descriptor for multi-view human action recognition from 4-D spatio-temporal interest points. Thus, we propose a hierarchical classification of the human activity recognition methods, which is depicted in Figure 10. We also presented multimodal approaches for the analysis of human social behaviors and interactions. A typical example of a behavior recognition system is depicted in Figure 10. "Show and tell: a neural image caption generator," in Proc. Tables 2 and 3 provide a comprehensive comparison of unimodal and multimodal methods, respectively, and list the benefits and limitations of each family of methods. IEEE International Conference on Computer Vision (Barcelona), 2556–2563. Song et al. doi:10.1109/TIP.2012.2197008 CrossRef Full Text | Google Scholar Yu, G., and Yuan, J. Google Scholar Yan, X., Kakadiaris, I. The problem of articulated 3D human pose estimation was studied by Fergie and Galata (2013), where the limitation of the mapping from the image feature space to the pose space was addressed using mixtures of Gaussian processes, particle filtering, and annealing (Sedai et al., 2013b). 14, 199–222. doi:10.1109/TITS.2013.2258152 CrossRef Full Text | Google Scholar Zhang, Z., Wang, C., Xiao, B., and Liu, S. Human behavior recognition in data-scarce domains. 73, 428–440. (2012a). Seo and Milanfar (2011) proposed a method based on space-time locally adaptive regression kernels and the matrix cosine measure. Signal Process. A fine separation between the meanings of "action" and "activity" was proposed by Turaga et al. Castellano et al. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1362–1369. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. Also, the number as well as the type of human activity classes to be recognized is an important factor that plays a crucial role in the robustness of the system. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (San Francisco, CA), 2061–2068. Vo et al. Multimodal cues are usually correlated in time, thus a temporal association of the underlying event and the different modalities is an important issue for understanding the data. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Anchorage, AK), 1–8. "Parsing videos of actions with segmental grammars," in Proc. Ni et al. Song et al. (2012) modeled social interactions by estimating the location and orientation of the faces of persons taking part in a social event, computing a line of sight for each face. Multimedia 16, 1766–1778. 48, 2377–2393. This feature fusion technique may improve recognition performance, but the new feature vector is of much larger dimension. International Conference on Machine Learning and Applications (Portland, OR), 240–245. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 1481–1488. doi:10.1016/j.cviu.2006.07.006 CrossRef Full Text | Google Scholar Ryoo, M. Most of the social networking systems that affect people's behavior, such as Facebook, Twitter, and YouTube, measure social interactions and infer how such sites may be involved in issues of identity, privacy, social capital, youth culture, and education. The final classification was performed by an SVM classifier. New York, NY: Springer-Verlag Inc. "Abnormal detection using interaction energy potentials," in Proc. M., and Narayanan, S. The whole approach was based on the construction of a space-time graph using a high-level descriptor, where the algorithm seeks to find the optimal subgraph that maximizes the activity classification score (i.e., find the maximum weight subgraph, which in the general case is an NP-complete problem). IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 740–747. (2014). IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1194–1201. Google Scholar Chen, L., Hoshi, T., and Schmid, C. Recently, a third approach to fusing the data has come to the foreground (Karpathy et al., 2014). Table 1. D., Cuellar, M. M., and Rao, B. (1976). Multimodal Methods Recently, much attention has been focused on multimodal activity recognition methods. 3, 184–198. Interactive phrases were treated as latent variables, while the recognition was performed using a CRF model. (2011a). 31, 203–221. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (San Francisco, CA), 2046–2053. UCF101: A Dataset of 101 Human Actions Classes from Videos in The Wild. Recognizing human actions using a new descriptor based on spatial-temporal interest points and weighted-output classifier. Yan et al. "Improved spatio-temporal salient feature detection for action recognition," in Proc. Figure 5 visualizes the decomposition of human activities according to their complexity. Mag. For example, video features are much more complex with higher dimensions than audio, and thus techniques for dimensionality reduction are useful. Expert Systems Appl. Their spatiotemporal relations were determined. Speech Communication. IEEE International Conference on Computer Vision (Graz), 428–441. "Learning realistic human actions from movies," in Proc. These features were tracked with respect to their velocities, and a generative mixture model was employed to learn the velocity history of these trajectories and classify each video clip. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 2855–2862. (2014) mixed image and motion features for online action classification. Y., and Kramer, G. K. 43, 1–43. doi:10.1016/j.patcog.2014.04.018 CrossRef Full Text | Google Scholar Gupta, A., Multimedia 15, 1553–1568. Activity-based person identification using fuzzy representation and discriminant learning. The authors described a variation of BoW method called bag-of-rectangles. doi:10.1109/TPAMI.2014.2303090 CrossRef Full Text | Google Scholar Kong, Y., Kit, D., and Fu, Y. Jung et al. "Multi-view pictorial structures for 3D human pose estimation," in Proc. 104, 90–126. (2014) exploited the interaction between human actions and scene geometry to recognize human activities from still images using 3D skeletal representation and adopting geometric representation constraints of the scenes. (2015) divided the human skeleton into five segments and used each of these parts to train a hierarchical neural network. The development of a fully automated human activity recognition system is a non-trivial task due to cluttered backgrounds, complex camera motion, large intraclass variations, and data acquisition issues. References Aggarwal, J. 29, 666–680. N., and Corso, J. "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in Proc. 104, 232–248. doi:10.1109/TPAMI.2013.2247823 CrossRef Full Text | Google Scholar Martinez, H. "Towards an affect-sensitive multimodal human–computer interaction," in Proc. Figure 12 illustrates the graphical models of different fusion approaches. Dollár et al. "Socially-aware large-scale crowd forecasting," in Proc. Comparison of multimodal methods. A non-linear knowledge transfer model based on deep learning was proposed for mapping action information from multiple camera views onto one single view. An overcomplete dictionary was constructed using a set of spatiotemporal descriptors. AlZoubi et al. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3193–3200. A complete survey, which covers important aspects of human activity recognition datasets, was introduced by Chaquet et al. Parliament dataset, which covers all the stances of politicians. Through another method, they introduced a novel behavior dataset, called the Parliament dataset, which consists of political speeches in the Greek parliament. The author used HMMs to model human actions as action units and then used grammatical rules to form a sequence of complex actions by combining different action units. doi:10.1016/j.jvcir.2013.03.008 CrossRef Full Text | Google Scholar Theodoridis, S., and Koutroumbas, K. Frequent changes in human motion and actions performed by groups of interacting persons make the problem amply challenging. Google Scholar Yu, G., Yuan, J., and Liu, Z. Predicting future events from partially unseen video clips with incomplete action execution has also been studied by Kong et al. Google Scholar Fothergill, S., Mentis, H. M., Nebel, J. The main problem of this work was the low resolution of the players to be tracked (a player was roughly 15 pixels tall). (1995). 38, 2437–2448. International Conference on Multimedia Interfaces (New York, NY), 239–248. 3. An audio-visual approach for recognizing dyadic interactions was presented by Yang et al. IEEE International Conference on Computer Vision (Sydney, NSW), 2280–2287. The individual motion patterns are analyzed separately and are then combined to estimate the event. Sigal et al. The temporal data were encoded into a trajectory space, which measures the similarity between activities and computes the angle between the associated subspaces. IEEE 90, 1151–1163. Affective computing studies model the ability of a person to express, recognize, and control his/her affective states in terms of hand gestures, facial expressions, physiological changes, speech, and activity recognition (Pantic and Rothkrantz, 2003). C., and Wang, X. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 1669–1676. Google Scholar Metallinou, A., Katsamanis, A., and Narayanan, S. The automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: a survey of related cues, databases, and tools. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 875–882. 1, 211–244. Pioneering this task, Wang and Mori (2008) were among the first to propose HCRFs for the problem of activity recognition. Finally, we provided the characteristics of building an ideal human activity recognition system. (2015) also divided human pose into smaller mid-level spatiotemporal parts. "Zero-shot recognition with unreliable attributes," in Proc. Intell. 10, 181–197. doi:10.1109/TPAMI.2013.137 CrossRef Full Text | Google Scholar Tran, K. Hidden part models for human action recognition: probabilistic versus max margin. The authors proposed a multitask learning approach Evgeniou and Pontil (2004) for simultaneously coping with low-level features and action-attribute relationships and introduced attribute regularization as a penalty term for handling irrelevant predictions. (2012). "Canal9: a database of political debates for analysis of social interactions," in Proc. 117, 1345–1355. A Mathematical Introduction to Robotic Manipulation, Edn. These attributes were treated as latent variables, which capture the degree of importance of each attribute for each action in a latent SVM approach. Google Scholar Song, Y., Morency, L. When the number of interacting people dynamically changes over time, the complexity of the problem increases and becomes more challenging. (2013a) proposed an analysis of multimodal data and datasets for spontaneous agreements and disagreements. F., Ellis, D. "Long-term recurrent convolutional networks for visual recognition and description," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 1725–1732. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1370–1377. Data fusion for visual tracking with particles. To capture these social interactions, eight subjects wearing head-mounted cameras participated in groups of interacting persons analyzing their activities from the first-person point of view. Most of the aforementioned methods do not require large training datasets, since they are able to model the hidden dynamics of the training data and incorporate prior knowledge over the representation of data. 46, 3223–3237. Human actions are treated as curves in the Lie group (Murray et al., 1994), and the classification was performed using SVM and temporal modeling approaches. "A spatio-temporal descriptor based on 3D-gradients," in Proc. doi:10.1162/153244301528126 CrossRef Full Text | Google Scholar Ntalianis, K., and Szegedy, C. M., Kohli, P., and Nowozin, S. P., van Balkegooij, A., de Jong, F., and Hiemstra, D. (2011c). (2014) focused on recognizing human actions that involve the manipulation of objects by representing a human action with a set of clustered motion trajectories. Given one training sample per class, the method can efficiently recognize this activity. Random forests for action representation have also attracted widespread interest for action recognition Mikolajczyk and Uemura (2008) and Yao et al. Rahmani et al. Google Scholar Shivappa, S., Trivedi, M. 6311 (Heronissos, Heraclion, Crete, greece: Springer), 494–507. Google Scholar Dalal, N., and Triggs, B. "Audio-visual emotion recognition using Gaussian mixture models for face and voice," in Proc. (2009). Trajectory-based activity recognition using: unsupervised, multilevel, and long-term adaptive approach. Audio-visual representation of human actions has gained an important role in human behavior recognition methods. doi:10.1016/j.imavis.2009.05.010 CrossRef Full Text | Google Scholar Oliver, N. (2015). Several issues concerning the minimum number of learning examples for modeling the dynamics of each class or safely inferring the performed activity label are still open and need further investigation. (2006). "Recording affect in the field: towards methods and metrics for improving ground truth labels," in Proc. This problem is considered as a part of the action recognition process. Action recognition relies heavily on the obtained pose estimations. "A Hough transform-based voting framework for action recognition," in Proc. Google Scholar Shafer, G. Intermediate semantic features representation for recognizing unseen actions in human activity recognition was proposed (Wang and Mori, 2010). The authors trained several models with latent variables to recognize human actions. To overcome this problem the authors proposed a patch-aware model, which learned regions of interacting subjects at different patch levels. Furthermore, the low layer of how many frames are required to recognize an action was addressed by Schindler and Gool (2008). Although both methods yielded promising results, they did not consider any kind of explicit correlation and/or association between the different modalities. "Human action segmentation with hierarchical supervoxel consistency," in Proc. 2009 IEEE International Conference on Computer Vision (Beijing), 1395–1402. R., and Shah, M. Multimodal methods describe atomic actions or interactions that may correspond to affective states of a person with whom he/she communicates and depend on emotions and/or body movements. doi:10.1016/j.patrec.2013.09.015 CrossRef Full Text | Google Scholar Tran, K. V., Zou, W. doi:10.1016/j.imavis.2011.07.004 CrossRef Full Text | Google Scholar Lillo, I., Soto, A., and Niebles, J. Self organizing maps (SOM) (Kohonen et al., 2001) were introduced by Iosifidis et al. Google Scholar Tian, Y., Sukthankar, R., and Shah, M. At first, they estimated which were the most important events, and then combined the different features using an AND/OR graph structure. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Portland, OR), 819–826. Computer Vision: Models Learning and Inference. L., Guibas, L. "DevNet: a deep event network for multimedia event detection and evidence recounting," in Proc. Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition. 35, 1635–1648. Representative frames of the main human action classes for various datasets. Neurocomputing 151, 554–564. Many of the existing datasets for human activity recognition were recorded in controlled environments, with participant actors performing specific actions. 4. 2. Stochastic Methods In recent years, there has been a tremendous growth in the amount of computer vision research aimed at understanding human activity. However, for high-level activities, the use of language priors is often not adequate, thus simpler and more explicit rules should be constructed. "Towards understanding action recognition," in Proc. T., Gao, B., Fei-Fei, L., and Koller, D. 13, 75–86. Human attributes from 3D pose tracking. Google Scholar Yacoob, Y., and Zelnik, Z. (2015) learned to arrange human actions in chronological order in an unsupervised manner by exploiting temporal ordering in video sequences. P., and Little, J. Google Scholar Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. This descriptor was a binary motion relationship descriptor for recognizing complex human interactions. (2013) employed spatiotemporal volumes using a deformable part model to train an SVM classifier for recognizing sport activities. Due to the large spatiotemporal redundancy in videos, many candidates may overlap. British Machine Vision Conference (Leeds: University of Leeds), 995–1004. "Learning semantic relationships for better action retrieval in images," in Proc. Google Scholar Fouhey, D. F., Delaitre, V., Gupta, A., Efros, A. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 5378–5387. The interacting groups were found by graph clustering, where each maximal clique corresponds to an interacting group. Gorelick et al. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 3686–3693. Neurocomputing 72, 39–46. Google Scholar Baxter, R. (2014b). The classification of a video sequence using local features in a spatiotemporal environment has also been given much focus. Audio and visual detection and segmentation were performed to extract the exact segments of interest in a video sequence, and then the influence model was employed. Complex activities are more difficult to recognize since space-time feature based approaches are employed. L., Tang, J. Conditional random fields, on the other hand, overcome the label bias problem. International Conference on Affective Computing and Intelligent Interaction (Memphis, TN), 107–116. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 2568–2577. To this end, Microsoft Kinect has played a significant role in motion capture of articulated body skeletons using depth sensors. doi:10.1109/TPAMI.2012.24 PubMed Abstract | CrossRef Full Text | Google Scholar Perez, P., Vermaak, J., and Blake, A. Google Scholar Donahue, J., Hendricks, L. This fact opens a promising research area that should be further studied. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., et al. An ideal action dataset should cover several topics, including diversity in human poses for the same action, a wide range of ground truth labels, and variations in image capturing and quality. Google Scholar Soleymani, M., Pantic, M., and Pun, T. Although most research focuses on leveraging human activity recognition from big data, this problem is still in its infancy. EURASIP J. Google Scholar Schuldt, C., Laptev, I., and Caputo, B. doi:10.1016/j.patcog.2004.12.013 CrossRef Full Text | Google Scholar Shen, Y., Chen, M., and Hauptmann, A. Google Scholar Moslund, T. Following a similar approach, Amiridi et al. Furthermore, viewpoint invariance is another issue that these approaches have difficulty in handling. Their feature vector was invariant to translation and was able to capture the relation between human body parts. Google Scholar Rodriguez, N. (2013). Seal et al. The authors modeled the relation of consecutive activities by building a graphical model for unsupervised learning of the activity label from depth sensor data. Intell. British Machine Vision Conference (Dundee), 1–12. Discriminative fusion of shape and appearance features for human pose estimation. 5.1. Affective Methods The core of emotional intelligence is understanding the mapping between a person's affective states and the corresponding activities, which are strongly related to the emotional state and communication of a person with other people (Picard, 1997). Sparse B-spline polynomial descriptors for human activity recognition. 3, 211–223. Thus, estimation of the maximum over variables using the Viterbi algorithm may be used. Andriluka and Sigal (2012) used a deformable part model as a kernel algorithm for estimating the most informative audio features. Google Scholar Wang, H., Kläser, A., Schmid, C., and Liu, C. doi:10.3390/s120201702 CrossRef Full Text | Google Scholar Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., et al. Google Scholar Kim, Y., Ayazoglu, M., Mao, T., Camps, O. Each action was described by a feature vector, which combines information about position, velocity, and local descriptors. Chen et al. Rule-based methods use a set of rules to describe human activities (Morariu and Davis, 2011; Chen and Grauman, 2012). (2014) trained a random decision forest (RDF) (Ho, 1995) and applied a joint representation of depth information and 3D skeletal positions for identifying human actions in real time. (2015) proposed a method for modeling the temporal structure of human actions that does not learn the temporal structure of human actions but rather decomposes human actions and scenes into the second category of methods, which is known as late fusion or fusion at decision level, combines several probabilistic models to learn the parameters of each modality separately. An invariant in translation and scaling descriptor was introduced by Oikonomopoulos et al. Google Scholar Jainy, M., Gemert, J. Also, all possible activities should be described during the recognition process, the recognition accuracy should be independent from the number of activity classes, and the activity identification process should be performed in real time and provide a high success rate and low false positive rate. Robust relative attributes for human action recognition. 2, 92–105. "Event recognition in videos by learning from heterogeneous web sources," in Proc. Continuous action recognition based on sequence alignment. IEEE Comput. K. Google Scholar Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Unimodal methods represent human activities from data of a single modality, such as images, and they are further categorized as: (i) space-time, (ii) stochastic, (iii) rule-based, and (iv) shape-based methods. P., Bengio, Y., and Yannakakis, G. "Modeling video evolution for action recognition," in Proc. K., and Ryoo, M. Cambridge, MA: MIT Press. (2014) leveraged the problem of human tracking for modeling the repulsion, attraction, and non-interaction effects in social interactions. In addition, annotating behavioral roles is time consuming and requires knowledge of the specific event. 34, 2441–2453. "The language of actions: recovering the syntax and semantics of goal-directed human activities," in Proc. Google Scholar Oikonomopoulos, A., Pantic, M., and Patras, I. Video Technol. This research area is generally considered to be a combination of computer vision, pattern recognition, artificial intelligence, psychology, and cognitive science. 6, 553–565. To solve the problem of body part correspondence between different views, the authors proposed a 3D pictorial structure representation based on a CRF model. Don't classify ratings of affect; rank them! IEEE Trans. Observing human-object interactions: using spatial and functional compatibility for recognition. Social network of interacting persons. "What do 15,000 object categories tell us about classifying and localizing actions?," in Proc. 29, 2247–2253. M. Moreover, social interactions are usually represented with complex temporal interaction between individuals. "Multiple granularity analysis for fine-grained action detection," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Miami Beach, FL), 1932–1939. doi:10.1016/j.patrec.2013.02.006 CrossRef Full Text | Google Scholar Chen, W., Xiong, C., Xu, R., and Corso, J. ACM International Conference on Multimedia Interaction (Santa Monica, CA), 27–30. Sapienza et al. Human activity recognition datasets. Comput. E., and Mitchell, T. A Gaussian process guided particle filter for tracking 3D human pose in video. 36, 1775–1788. Zhou and Wang (2012) proposed a new representation of local spatiotemporal cuboids for action recognition. The 2D poses for different sources were coupled into 3D space using a mixture of multiview pictorial structures models. "Action recognition in cluttered dynamic scenes using pose-specific part models," in Proc. The recognition accuracy of such complex videos can also be improved by relating textual descriptions and visual context to a unified framework (Ramanathan et al., 2013). (2014) have also employed sparse representation of skeletal data in the dissimilarity space for human activity recognition. Relevant information was summarized together through a ranking framework. Gestures are considered as primitive movements of the body parts of a person that may correspond to a particular action of this person (Yang et al., 2013). doi:10.1007/s11263-014-0758-9 CrossRef Full Text | Google Scholar Kulkarni, K., Sharma, G., Zepeda, J., and Chevallier, L. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Portland, OR), 2642–2649. (2013). doi:10.1007/s00371-014-1005-8 CrossRef Full Text | Google Scholar Nicolaou, M. Sargin et al. T., et al. In such cases, the bag of words (BoW) approach was employed. "Learning to recognize human actions from videos in three sources via fine-grained multimodal features," in Proc. "How do we interact when you're looking?" (2013) proposed an unsupervised method for recognizing motion primitives for human action classification from a set of very few examples. J., Nou Salinas, R. Multimedia 16, 289–298. Google Scholar Wang, Y., Guan, L., and Venetsanopoulos, A. Google Scholar Patron-Perez, A., Marszalek, M., Reid, I., and Zisserman, A. (2015) mined appearance and motion features for recognizing group activities in crowded scenes collected from the web. Their system was able to preserve non-linear relationships between multimodal features and showed that unsupervised learning can be used efficiently for feature selection. 43, 875–885. doi:10.1016/j.imavis.2012.08.018 CrossRef Full Text | Google Scholar Metallinou, A., and Narayanan, S. "Action recognition by exploring data distribution and feature correlation," in Proc. Pei et al. (2014) decomposed the problem of complex activity recognition into two sequential sub-tasks with increasing granularity levels. doi:10.1109/TNNLS.2011.2181865 CrossRef Full Text | Google Scholar Jaimes, A., and Sebe, N. It is evident that the lack of large and realistic human activity recognition datasets is a significant challenge that needs to be addressed. "Classifying behavioral attributes using conditional random fields," in Proc. 4.1. Space-Time Methods Space-time approaches focus on recognizing activities based on space-time features or on trajectory matching. (B) Examples of correct human pose estimation of complex activities. Finally, events are high-level activities that describe social actions between individuals and indicate the intention or the social role of a person (Lan et al., 2012a). "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in Proc. This limitation has resulted in a lack of attention on building a graphical model for human activity recognition. Furthermore, the authors were able to recognize action units that exist in real actions, since the human actions are defined to be of the same class. "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in Proc. More attention should be put in developing robust methods under the uncertainty of missing data either on training steps or testing steps. "Discriminative figure-centric models for joint action localization and recognition," in Proc. C., Escorcia, V., Ghanem, B., and Niebles, J. Lui et al. As a human silhouette consists of limbs jointly connected to each other, it is important to obtain exact human body parts from videos. A probabilistic multimodal approach for emotion recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (Vancouver, BC), 3687–3691. doi:10.1016/j.cviu.2009.08.001 CrossRef Full Text | Google Scholar Lafferty, J. Another challenge worthy of further exploration is the exploitation of unsupervised sequences, where one activity may succeed another. "Real time action recognition using histograms of depth gradients and random decision forests," in Proc. However, natural video sequences may contain irrelevant scenes or scenes with multiple actions. Finally, the third approach is a combination of the above two. Google Scholar Iosifidis, A., Tefas, A., and Pitas, I. M., Xiang, T., and Gong, S. G., Ye, G., Chang, S. The BoW technique was used to form a vocabulary of human actions, whereas agglomerative information bottleneck and SVM were used for action classification. Google Scholar Yu, Z., Deng, F., and Pantic, M. As a result of this research, many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. Human location was treated as a latent variable, which was extracted from a discriminative latent variable model by simultaneous recognition of an action. 23, 1310–1323. L., and Davis, L. Researchers have conceived and used many stochastic techniques, such as hidden Markov model (HMMs) (Bishop, 2006) and hidden conditional random fields (HCRFs) (Quattoni et al., 2007), to infer useful results for human activity recognition. On the other hand, more complex activities such as "peeling an apple," are more difficult to identify. Humans have the ability to understand another human's actions by interpreting stimuli from the surroundings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fourth Edn. The combination of visual features with other more informative features, which reflect human emotions and psychology, is necessary for this task. The problem of appearance-to-pose mapping for human activity understanding was studied by Urtasun and Darrell (2008). (2014) presented a method for classifying the classification of affective and cognitive states of computer game players. Election. Figure 11 shows the resulting social network built from this method. Google Scholar Iküzler-Cinbis, N., and Duygulu, P. Overcoming this limitation may lead to better results. CVPR. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Stochastic modeling of human activities is in an active research area. Google Scholar Chen, L., Wei, H., and Ferryman, J. doi:10.1109/TSMCA.2012.2226575 CrossRef Full Text | Google Scholar Wu, Q., Wang, Z., Deng, F., and Feng, D. Google Scholar Chen, C. Grzegorzek, C. K., Chellappa, R., Subrahmanian, V. Google Scholar Pirsiavash, H., and Ramanan, D. 2 (Nice), 726–733. Google Scholar Khamis, S., Morariu, V. "Associating textual features with visual ones for improving video action recognition," in Proc. Vo et al. Using the DCT of human activities, each view was modeled as a sub-feature set of a human body joint, while the classification of human actions were introduced by Wu et al. "Learning to track and identify players from broadcast sports videos," in Proc. ... Stochastic methods have, and Xia, L., Niebles et al. Example of a typical human space-time method based on dense trajectories (bottom row). HMMs treat features as conditionally independent, but this assumption may not hold for the majority of applications. The work of Vrigkas et al. European Conference on Computer Vision (Florence), 693–706. 116, 396–410. "Recognizing actions by shape-motion prototype trees," in Proc. Iküzler-Cinbis and Sclaroff (2010) extracted dense features and performed tracking over consecutive frames for describing both motion and shape information. Google Scholar Le, Q. (2006). ACM Computer Google Scholar Wu, Q., Wang, Z., Deng, F., Chi, Z., and Feng, D. This dataset should be sufficiently rich in a variety of human actions. Action units were summarized together in a ranking framework. "Learning to recognize human activities in video," in Proc. (2014) employed a set of text descriptions for recognizing activities in text videos. Google Scholar Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. "Elastic functional coding of human actions: from vector-fields to latent variables," in Proc. However, such information is not sufficient to fully understand the underlying activity as it does not capture the variations in human poses when interacting with other persons. To overcome these problems, a task is required that consists of three components, namely: (i) background subtraction (Elgammal et al., 2002; Mumtaz et al., 2014), in which the system attempts to separate the parts of the image that are invariant over time (background) from the objects that are moving or changing (foreground); (ii) human tracking, in which the system locates human motion over time (Liu et al., 2010; Wang et al., 2013; Yan et al., 2014); and (iii) human action and object detection (Pirsiavash and Ramanan, 2012; Gan et al., 2015; Jainy et al., 2015), in which the system is able to localize a human activity in an image. The problem of identifying multiple persons simultaneously and performing action recognition was presented by Khamis et al. Structured learning of human interactions in TV shows. More sophisticated high-level activity recognition methods need to be developed, which should be able to recognize simultaneously occurring activities. doi:10.1007/s11263-013-0677-1 CrossRef Full Text | Google Scholar Gao, C., Wang, W., Yang, Y., Yuan, Y., Yang, Y. 12, 90–103, where a sparse representation was used to model human activity classes. F., Yang, W., Wang, Y., and Mori, G. H., Robertson, N. J., and Fei-Fei, L. Google Scholar Wang, Z., Wang, J., Xiao, J., Lin, K. A key challenge in recognizing human behaviors is to define specific emotional attributes for multimodal dyadic interactions (Metallinou and Narayanan, 2013). They used an agglomerative approach to merge super-voxels that share common attributes and localize human activities. A new social activity dataset has also been proposed. Image Understand. D., and Vidal, R. (2010). Learn. Google Scholar Sigal, L., Isard, M., Haussecker, H. 4738 (Lisbon), 71–82. doi:10.1007/s11263-012-0594-8 CrossRef Full Text | Google Scholar Yang, Y., Saleemi, I., Liu, C., Ilc, Y., and Yuan, J. "The USC creative IT database: a multimodal database of theatrical improvisation," in Proc. (2007) considered actions as 3D space-time silhouettes of moving humans. Google Scholar Thurau, C. Both architectures and unsupervised feature learning have garnered considerable recent research interest (Ngiam et al., 2011). (2006) studied several state-of-the-art methods of human behavior recognition including affective and social cues and covered many open computational problems and how they can be efficiently incorporated into a human–computer interaction system. A., Laptev, I., and Sivic, J. Ratings are one of the most popular affect annotation tools. The easiest way to gain the benefits of multiple features is to directly concatenate features in a larger feature vector and then learn the underlying action (Sun et al., 2009). Circuits Syst. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 2361–2368. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3153–3160. A significant problem in constructing a proper human activity recognition dataset is the annotation of each action, which is generally performed manually by the user, making the task biased. Emotions, facial expressions, shoulder gestures, and audio cues are combined for continuous prediction of emotional states (Nicolaou et al., 2011). This correlation characterizes the group of persons P1 ... P25 and the subjects wearing the cameras S1 ... S8 are weighted based on how often a person's face is captured by a subject's camera (Fathi et al., 2012). 45, 1–33. Google Scholar Yang, Y., and Mori, G. The human ability to recognize another person's activities is one of the main subjects of study of the scientific areas of computer vision and machine learning. Boston: Academic Press. (2013b) and Ye et al. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (Shanghai), 1–8. The main disadvantage of using a global representation, such as optical flow, is the sensitivity to noise and partial occlusions. doi:10.1109/TNNLS.2014.2222488 CrossRef Full Text | Google Scholar Rodriguez, M., Nowozin, S., Lampert, C., and Li, F. Google Scholar Morency, L., and Pantic, M. An attribute-based action recognition method was presented by Marszalek et al. However, their work cannot deal with actions in which only the upper body part is visible. The rest of the paper is organized as follows: in Section 2, a brief review of previous surveys is presented. A facet model for image data. Sundaram, Kohanyi S., Babkhami and Leong (2013) also proposed a hierarchical representation of videos segments for recognizing atomic actions by building a codebook of spatiotemporal volumes. The BoW technique was used to form a vocabulary of human actions, whereas agglomerative information bottleneck and SVM were used for action classification. IEEE Computer Society Conference on Computer Vision (Sydney, NSW), 905–912. Probabilistic approach to detecting dependencies between data sets. Trajectory-based methods face the problem of human body detection and tracking, as these are still open issues. A survey of affect recognition methods: audio, visual, and spontaneous expressions. 22, 26–27. Sel. To avoid overfitting, they proposed a novel classification technique combining Adaboost and sparse representation algorithms. "Social behavior recognition in continuous video," in Proc. (2011) presented a boosting algorithm called LatentBoost. Sun et al. A combination of discriminative and generative models improved the estimation of human pose. Image Represent. Instead of using bounding box traces for human actions classification, Jainy et al. "ACTIVE: activity concept transitions in video event classification," in Proc. Early fusion, or fusion at the feature level, combines features of different modalities, usually by reducing the dimensionality in each modality and creating a new feature vector that represents an individual. C., and Snoek, C. "Deep learning for robust feature generation in audiovisual emotion recognition," in Proc. Salt Lake City: A Human Face. Extending the previous method, Gupta et al. Nevertheless, identifying which body parts are most significant for recognizing complex human activities still remains a challenging task (Lillo et al., 2014). The authors took advantage of facial expressions as they can be expressed by the facial action coding system (FACS) (Ekman et al., 2002), which describes all possible facial expressions as a combination of action units (AU), and combines them with audio information, extracted from each participant, to identify their emotional state. doi:10.1145/2523819 CrossRef Full Text | Google Scholar Rohrbach, M., Amin, S., Myklaykh, A., and Schiele, B. A Gaussian mixture model was used to cluster the motion trajectories, and the action labeling was performed using a nearest neighbor classification scheme. A survey on vision-based human action recognition. (2012b) recognized group activities, which were considered as latent variables, encoding the contextual information in a video sequence. Siddiquie et al. doi:10.1016/j.cviu.2006.08.002 CrossRef Full Text | Google Scholar Morariu, V. Human activity as a manifold-valued random process. Online action recognition using covariance of shape and motion. 36, 1299–1311. Many works on human activity recognition based on deep learning techniques have been proposed in the literature. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1322–1329. Effective tracking of human body using a manifold hierarchy. Google Scholar Shao, J., Kang, K., Loy, C. Also, Yang et al. (2013) used dense optical flow trajectories to describe the kinematics of motion patterns in video sequences. doi:10.1016/j.neucom.2012.02.002 CrossRef Full Text | Google Scholar Zhang, Y., and Mori, G. Similar in spirit to the previous works, Yang et al. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 812–819. Z. We discussed unimodal approaches and provided an internal categorization of these methods, which were developed for analyzing gesture, atomic actions, and more complex activities, either directly or employing activity decomposition into simpler actions. Modeling 3D data is also a new trend, and it was extensively studied by Chen et al. Ni et al. "ActivityNet: a large-scale video benchmark for human activity understanding," in Proc. Most of the existing works do not take into account the fact that a specific action may be performed in different manner by a different actor. Table 1 summarizes the previous surveys on human activity and behavior recognition methods sorted by chronological order. S., and Yang, Y. Multimedia 16, 1–10. Google Scholar Vrigkas, M., Nikou, C., and Kakadiaris, I. Google Scholar Wang, J. ... Pattern Recognit. Each frame was segmented into body parts that correspond to a specific spatiotemporal interest point. Human Action Recognition Each video was segmented into periods of activities by constructing a temporal context that combined local and global representations of human actions. Lin and Ni. "Video summarization by learning submodular mixtures of objectives," in Proc. Google Scholar Schmidt, A. Zhang et al. Human activities from unsegmented videos using a voting-based classification scheme to find the most frequently used action label. As a result, Bandla and Grauman (2013) proposed a method for recognizing human actions. (2013) introduced a representation of human actions, called the "poselet activation vector," which is defined by the 3D orientation of the head and torso and provided a robust representation of human pose and appearance. Google Scholar Pantic, M., and Rothkrantz, L. 31, 39–58. doi:10.1007/s11263-007-0122-4 CrossRef Full Text | Google Scholar Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C., Lee, J. 6, 396, 404–416. (B) Hierarchical latent discriminative model proposed by Song et al. Human activities were represented by bag-of-video words constructed from spatial temporal interest points by extracting motion and appearance features and building a vocabulary of visual words. A human activity was considered as a sequence of shapes, which is considered as one realization of a random process on a manifold. Classification of actions from human poses (Lan et al., 2014). They constructed 3D body part hypotheses by triangulation of 2D pose detections. However, recognizing human actions from still images by taking advantage of contextual information, such as surrounding objects, is a very active topic (Yao and Fei-Fei, 2010). In addition, audio-visual analysis is used in many applications not only for audio-visual synchronization (Lichtenauer et al., 2011) but also for tracking (Perez et al., 2004) and activity recognition (Wu et al., 2010). Human action was treated as a latent variable in an HMM model, while a Fisher kernel technique (Perronnin and Dance, 2007) was employed to model human activity with a four-length temporal structure. Google Scholar Amin, S., Andriluka, M., Rohrbach, M., and Schiele, B. Combining both pose-specific appearance and the joint appearance of body parts helps to construct a more powerful representation of the human body. 16, 345–379. Figure 3. Each node represents one person and each edge on the graph is associated with a weight according to the level of the interaction between the participants. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3185–3192. Google Scholar Gaidon, A., Harchaoui, Z., and Schmid, C. Google Scholar Sigal, L., Isard, M., Haussecker, H. 3D human pose estimation and action recognition. (2013) employed spatiotemporal volumes using a deformable part model to train an SVM classifier for recognizing sport activities. The interaction between the first-person and the observed people was modeled using a first-person-centric graphical model. Google Scholar Yi, Y., and Lin, M. (2014) proposed a method for detecting human actions by combining spatial and temporal features. To this end, it is considered as a social activity dataset. ... IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3185–3192. Google Scholar Bishop, C. The goal of human activity recognition is to examine activities from video sequences or still images. Human behavior recognition in data-scarce domains. The proposed method was tested on datasets that were annotated in the video, which was able to recognize human actions in realistic video sequences. (2015) proposed a new method for human activity recognition from still images. International Journal of Computer Vision. Multimodal Methods ... An overview of the types of the activities where a fusion technique may be employed to recognize human actions and interactions. Google Scholar Haykin, S., and Ng, A. Attributes correspond to general prior knowledge, which may be considered as a stochastically predictable sequence of states. All features were projected onto a common subspace, and a boosting technique was employed to recognize human actions from labeled and unlabeled data. doi:10.1007/s10044-013-0349-3 CrossRef Full Text | Google Scholar Zhou, Q., and Wang, G. International Conference on Articulated Motion and Deformable Objects (Mallorca), 114–123. CoRR, abs/1212.0402. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 1293–1301. The extracted features were more robust under geometric transformations than the features described by a Gabor filter (Fogel and Sagi, 1989). Foulhey et al. Ma et al. Recasens, N. J. Springer. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 3337–3344. Video event understanding using natural language descriptions. Google Scholar Burbury, D. The authors introduced the "deep network model" in which the human activities are modeled by using the grammar rules. Human activity recognition in a multimodal perspective. The authors in video... (2014) employed the RGB-D sensor, which is able to provide a set of skeletal data in depth images, to understand the interactions between human body parts and objects. The task is considered as a problem of recognizing the emotional states of dyadic interactions. Man Cybern. However, this method depends on the initialization of the random walk process. However, they are usually more complicated than non-parametric methods, since they use dynamic programming or computationally expensive HMMs for estimating a varying number of parameters. International Conference on Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science, Vol. doi:10.1016/j.cviu.1998.0716 CrossRef Full Text | Google Scholar Rahmani, H., Mahmood, A., Huynh, D., and Mian, A. (2014) proposed an "action-pose" model that combines pose estimation and action recognition, and introduced a new dataset. Google Scholar Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. IEEE Conference on Computer Vision and Pattern Recognition (Boston, MA), 961–970. In this context, our goal is to ... "Beyond actions: discriminative models for contextual group activities," in Proc. P., and Sebe, N. "Cross-view action modeling, learning, and recognition," in Proc. A fundamental problem in behavior recognition is the human pose estimation. The authors also proposed a new model that learns "actionlet ensemble model" which captured the intraclass variations and was robust to errors incurred by depth cameras. K. Hossain, M. Annual Conference on Neural Information Processing Systems (Vancouver, BC), 1410–1418. "Recognizing group activities using poselet-based features," in Proc. The problem of identifying and modeling multiple activities in a video sequence was reported by Zhang et al. Satkin and Hebert (2010) explored the effectiveness of video segmentation by discovering the most significant portions of videos. Boca Raton, FL: CRC Press, Inc. Users cannot always express their emotion with words, and producing satisfactory and reliable ground truth that corresponds to a given training instance is quite difficult as it can lead to ambiguous and subjective labels. (2012) recognized human actions using stereo cameras. O. Multi-view gesture recognition using Gaussian mixture models (GMM) (Bishop, 2006) to effectively capture human actions in the temporal domain while keeping the features invariant to the location of the interest points at several space and time scales. (2004). Google Scholar Xu, R., Agarwal, P., Kumar, S., Krovi, V. Multimedia 15, 1396–1403. A great difficulty in multimodal feature analysis is the dimensionality of the data from different modalities. International Journal on Visual Emotion Challenge and Workshop, Lecture Notes in Computer Science, Vol. A recent survey on human behavior recognition provides a complete summarization of up-to-date techniques for automatic human behavior analysis for single person, multiperson, and object–person interactions (Candamo et al., 2010). W., and Davis, L. On the other hand, the accuracy of the temporal correlations ignores the domination of one modality over the others. Then all scores are combined together by a transferred ranking yielding a final action score (Westerveld et al., 2003; Jiang et al., 2014). Recognizing 50 human action categories of web videos. Audiovisual synchronization and fusion using canonical correlation analysis. In particular, we may conclude that despite the tremendous increase of human understanding methods, many problems still remain open, including modeling of human poses, handling occlusions, and annotating data. Figure 5. Graph. To this end, several 2D human pose benchmarks have been proposed for the evaluation of articulated human pose estimation methods (Andriluka et al., 2014). 110, 259–274. Among various classification techniques, two main questions arise: "What action?" (i.e., the localization problem). "What action?" (i.e., the recognition problem) and "Where in the video?" (i.e., the localization problem). Ji, Tran, C., Trivedi, M. A separate classifier for each source is learned and a multidomain adaptation approach was followed to infer the labels for each data source. Similar in spirit is the work of Wu et al. Google Scholar Roshtkhari, M. Google Scholar Gao, Z., Zhang, H., Xu, G. Behavioral methods aim to recognize behavioral attributes, non-verbal multimodal cues, such as gestures, facial expressions, and auditory cues (Song et al., 2012a; Vrigkas et al., 2014b). (2008) addressed the problem of action recognition by creating a codebook of space-time interest points. (2013) analyzed four different affective dimensions, such as activation, expectancy, power, and valence (Schuller et al., 2011). "Histograms of oriented gradients for human detection," in Proc. Turaga, P. Advances in Neural Information Processing Systems (Montreal, QC), 3464–3472. The classification model and some representative examples of the estimation of human pose are depicted in Figure 8. Vis. Google Scholar Burenius, M., Sullivan, J., and Carlsson, S. 5.4. Multimodal Feature Fusion Consider the scenario where several people have a specific activity/behavior and some of them may emit sounds. "Human detection using oriented histograms of flow and appearance," in Proc. Figure 8. (A) Graphical model. Google Scholar Ikizler, C. International Conference on Articulated Motion and Deformable Objects (Mallorca), 260–272. Google Scholar Jain, M., Gemert, J., Jegou, H., Bouthemy, P., and Snoek, C. 7, 5, 314–328. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1274–1281. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 1226–1233. The exact application (i.e., learning human activities from very little training data or missing data) is also very challenging. British Machine Vision Conference (Dundee), 1–11. "Cross-view activity recognition using hankelets," in Proc. "2D human pose estimation: new benchmark and state of the art analysis," in Proc. Google Scholar Ni, B., Paramathayalan, V. On the other hand, human activities that correspond to fewer data may not perform well. Ramanathan et al. Mixtures of Gaussian process models for human pose estimation. Nonetheless, the description of human activities with high-level contents usually leads to recognition methods with high computational complexity. The work of Sanchez-Riera et al. K., Sonhak, S. Furthermore, their method was later fused to combine audio and visual information together. Social interaction can be considered as a special type of activity where someone adapts his/her behavior according to the group of people surrounding him/her. Self-Organizing Maps, Third Edn. Figure 1. Google Scholar Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. First, dense feature sampling is performed for capturing local motion. "A large-scale benchmark dataset for event recognition in surveillance video," in Proc. P., and Pantic, M. In the sense of video labeling, the study of Wang et al. P., Lilius, J., and Calvo-Flores, M. Each of these sub-categories describes specific attributes of human activity recognition methods according to the type of representation each method uses. M., Carmona, E. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 368–375. H., and Huang, T. For example, a soccer player interacts with a ball when playing soccer. Modeling crowded scenes has been a difficult task due to partial occlusions, interacting motion patterns, and sparsely distributed cameras in outdoor environments (Alahi et al., 2014). IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 961–970. "Identifying players in broadcast sports videos using conditional random fields," in Proc. S., Calvo, R. "Discriminative hierarchical modeling of spatio-temporally composable human activities," in Proc. The main disadvantage of this method is that it used different classifiers to separately learn the audio and visual context. Du et al. Google Scholar Yang, Y., Saleemi, I., and Shah, M. (2007) suggested a method for speaker identification integrating a hybrid scheme of early and late fusion of audio-visual features and used CCA (Hardoon et al., 2004) to synchronize the multimodal features. Although a list of action datasets that correspond to most of these specifications has been introduced in the literature, the question of how many actions we can actually learn is a task for further exploration. Yu and Yuan (2015) extracted bounding box candidates from video sequences, where each candidate may contain human motion. doi:10.1006/cviu.1998.0744 CrossRef Full Text | Google Scholar Aggarwal, J. C. The different types of descriptors were fused at the decision level using a discriminative learning model. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 4362–4370. L. They incorporated psychological signals into emotional states, such as relaxation, anxiety, excitement, and fun, and demonstrated that deep learning was able to extract more informative features than feature extraction on psychological signals. Some approaches use snippets of motion trajectories (Matikainen et al., 2009; Raptis et al., 2012), while others use the full length of motion curves by tracking the optical flow features (Vrigkas et al., 2014a). Each activity is considered as a set of primitive rules/attributes, which enables the construction of a descriptive model for human activity recognition. Annual ACM International Conference on Multimedia (Singapore), 399–402. British Machine Vision Conference (Bristol), 1–12. (eds) (2001). B., Hilton, A., and Krüger, V. doi:10.1007/s00530-010-0182-0 CrossRef Full Text | Google Scholar Bandla, S., and Grauman, K. IEEE International Conference on Computer Vision (Sydney, NSW), 2712–2719. The type and amount of data that each approach uses depends on the ability of the underlying algorithm to deal with heterogeneous and/or large scale data. A novel method for fusing verbal (i.e., textual information) and non-verbal (i.e., visual signals) cues was proposed by Evangelopoulos et al. A person adapts his/her behavior according to the person with whom s/he interacts. A survey of video datasets for human action and activity recognition. P., and Davis, R. European Conference on Computer Vision (Heraklion), 536–548. 47, 1626–1641. Usually, the terms "activity" and "behavior" are used interchangeably in the literature (Castellano et al., 2007; Song et al., 2012a). Google Scholar Bousmalis, K., Mehu, M., and Pantic, M. "3D model-based continuous emotion recognition," in Proc. Tang et al. Google Scholar Ni, B., Moulin, P., Yang, X., and Yan, S. 34, 1691–1703. Cost-effective solution to synchronised audio-visual data capture using multiple sensors. S. One disadvantage of this method is that it cannot deal with self-occlusions (i.e., overlapping parts of human skeleton). B., and Kasturi, R. Moreover, the computation of these features produces sparse and varying numbers of detected interest points, which may lead to low repeatability. Most of these reviews summarize human activity recognition methods, without providing the strengths and the weaknesses of each category in a concise and informative way. (2015) were able to transfer semantic knowledge between classes to learn human actions from still images. "Trajectons: action recognition through the motion analysis of tracked features," in Workshop on Video-Oriented Object and Event Classification, in Conjunction with ICCV (Kyoto: IEEE), 514–521. Pantic et al. Some representative frames that summarize the main human action classes are depicted in Figure 3. However, the maximum set coverage problem is NP-hard, and thus the estimation requires approximate solutions. (2008), where the activity recognition methods were categorized according to their degree of activity complexity. (2013) studied the problem of heterogenous feature combination for recognizing complex events. The advantage of early fusion is that it yields good recognition results when the different modalities are highly correlated, since only one learning phase is required. Ikizler and Duygulu (2007) modeled the human body as a sequence of oriented rectangular patches. (2015) extracted spatiotemporal segments from video sequences that correspond to whole or part human motion and constructed a tree-structured vocabulary of similar actions. Google Scholar Dalal, N., Triggs, B., and Schmid, C. 45, 2562–2572. V. Google Scholar Smola, A. Figure 4 depicts an example of a space-time approach based on dense trajectories and motion descriptors (Wang et al., 2013). Based on the histograms of oriented Gaussians, Dalal and Triggs (2005) were able to detect humans, whereas classification of actions was made by training an SVM classifier. The generation of short description from video sequences (Vinyals et al., 2015) based on convolutional neural networks (CNN) (Ciresan et al., 2011) was also used for activity recognition (Donahue et al., 2015). Human activity analysis: a review. 33, 2287–2301. Thus, building strong models that can cope with multimodal data, such as gestures, facial expressions and psychological data, depends on the ability of the model to discover relations between different modalities and generate informative representation on affect annotations. Similar in spirit, the work of Jain et al. Pattern Recognition and Machine Learning. Specific attributes are predicted from already learned classifiers and are mapped into a class-level score. Spatially oriented histograms were formed to describe a human action, while the classification of an action was performed using four different methods, such as frame voting, global histogramming, SVM classification, and dynamic time warping (DTW) (Theodoridis and Koutroumbas, 2008). An ideal human activity dataset should address the following issues: (i) the input media should include either still images and/or video sequences, (ii) the amount of data should be sufficient, (iii) input media quality (resolution, grayscale or color), (iv) large number of subjects performing an action, (v) large number of action classes, (vi) changes in illuminations, (vii) large intraclass variations (i.e., variations in subjects' poses), (viii) photo shooting under partial occlusion of human structure, and (ix) complex backgrounds. Google Scholar Lan, T., Sigal, L., and Mori, G. Semi-supervised multiple feature analysis for action recognition. "Recognize complex events from static images by fusing deep channels," in Proc. Google Scholar Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. Kolb (Berlin Heidelberg: Springer), 149–187. (2012b) addressed the multiview human-tracking problem where the modeling of 3D human pose consisted of a collection of human body parts. Efros et al. However, the effectiveness of such methods is limited by tracking inaccuracies in human poses and complex backgrounds. (2014) proposed a hierarchical method for predicting future human actions, which may be considered as a reaction to a previous performed action. "Better exploiting motion for better action recognition," in Proc. European Conference on Computer Vision (Zurich), 29–44. Actions as space-time shapes. International Conference on Machine Learning (Williamstown, MA: Williams College), 282–289. Infinite hidden conditional random fields for human behavior analysis. Google Scholar Kong, Y., Jia, Y., and Fu, Y. 31, 137–152. In the literature, there are two main fusion strategies that can be used to tackle this problem (Atrey et al., 2010; Shivappa et al., 2010). 7575 (Florence), 530–543. Furthermore, it is evident that there exists a great need for efficiently manipulating training data that may come from heterogeneous sources. (2015) proposed to incorporate information from human-to-objects interactions and combined several datasets to transfer information from one dataset to another. Google Scholar Kohonen, T., Schroeder, M. Human behaviors refer to physical actions that are associated with the emotions, personality, and psychological state of the individual (Martinez et al., 2014). (2011) considered a system based on HCRFs for spontaneous agreement and disagreement recognition using audio and visual features. "Social interactions: a first-person perspective," in Proc. R. Liu et al. 18, 1473–1488. (2005) proposed spatiotemporal features based on cuboid descriptors. Therefore, a comprehensive evaluation of feature fusion methods that retain the feature coupling is an issue that needs to be assessed. Discriminative latent models for recognizing contextual group activities. (2012) investigated the properties of developing a user-independent emotion recognition system that is able to detect the most informative affective tags from electroencephalogram (EEG) signals, pupillary reflex, and bodily responses that correspond to video stimulus. These intermediate features were learned during training, while parameter sharing between classes was enabled by capturing the correlations between frequently occurring low-level features (Akata et al., 2013). Social network extraction and analysis based on multimodal dyadic interaction. Guha and Ward (2012) employed a technique of sparse representations for human activity recognition. 47, 3343–3361. doi:10.1016/j.neucom.2014.06.085 CrossRef Full Text | Google Scholar Gavrila, D. "Multimodal human-computer interaction: a survey," in Computer Vision and Image Understanding, Vol. Marín-Jiménez et al. "Regularized multi-task learning," in Proc. View-invariant action recognition based on artificial neural networks. Google Scholar Yao, B., and Fei-Fei, L. (2012) proposed a novel method applying surround suppression. Audio-visual features and emotional annotations are fed into a GMM for estimating the emotional curves (Metallinou et al., 2013). M., Divakaran, A., and Sawhney, H. The trajectories were clustered by an SVM classifier. Koch, and A. Space-time approaches can hardly recognize actions when more than one person is present in a scene. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Colorado Springs, CO), 1297–1304. "Strong appearance and expressive spatial models for human pose estimation," in Proc. Previous Surveys and Taxonomies There are several surveys in the human activity recognition literature. Figure 4. "Modeling supporting regions for close human interaction recognition," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 780–787. Deep learning has gained much attention for multisource human pose estimation (Ouyang et al., 2014) where the tasks of detection and estimation of human pose were jointly learned. (2011b) associated multimodal features (i.e., textual and visual) for classifying affective states in still images. N., and Mori, G. "Discriminative virtual views for cross-view action recognition," in Proc. Pattern Anal. Human activity recognition in videos using a single example. doi:10.1016/j.patrec.2014.04.011 CrossRef Full Text | Google Scholar Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. IEEE International Conference on Computer Vision (Rio de Janeiro), 1–8. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 3762–3771. "Hybrid fusion approach for detecting affects from multichannel physiology," in Proc. M., and Moeslund, T. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 2625–2634. A comparison of early versus late fusion methods for video analysis was reported by Snoek et al. The author combined a GMM with a Fisher kernel to model multimodal dyadic interactions and predict the body language of each subject according to the behavioral state of his/her interlocutor. P. (2011b) performed human activity recognition by associating the context between interest points based on the density of all features observed. "Pose primitive based human action recognition in videos or still images," in Proc. A video segmentation approach for video activities and a decomposition into smaller clips task that contained sub-actions was presented by Wu et al. D., Ahmed, J., and Shah, M. 8445 (Ioannina), 95–104. (2009) introduced spatial and functional constraints on static shape and appearance features and they were also able to identify human-to-object interactions without incorporating any motion information. Matching mixtures of curves for human action recognition. B., Moeslund, T. Most of the existing probabilistic methods for human activity recognition may perform well and apply exact and/or approximate learning and inference. Google Scholar Ferrari, V., Marin-Jimenez, M., and Zisserman, A. Image Vis. doi:10.1016/S0031-3203(02)00100-0 CrossRef Full Text | Google Scholar Wang, S., Ma, Z., Yang, Y., Li, X., Pang, C., and Hauptmann, A. The interaction between different classes was performed using linguistic rules. (A) Factorized HCRF model used by Wang and Mori (2008). The survey of Moeslund et al. A novel part-based skeletal representation for action recognition was introduced by Vemulapalli et al. Then, features are tracked using dense optical flow, and feature descriptors are computed (Wang et al., 2013). doi:10.1109/TPAMI.2007.70711 PubMed Abstract | CrossRef Full Text | Google Scholar Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R. Factors that can affect human behavior may be decomposed into several components, including emotions, moods, actions, and interactions, with other people. (2006) mainly focused on pose-based action recognition methods and proposed a fourfold taxonomy, including initialization of human motion, tracking, pose estimation, and recognition methods. "Social roles in hierarchical models for human activity recognition," in Proc. A., Murphy, K. Google Scholar Sedai, S., Bennamoun, M., and Huynh, D. It is necessary for the system to be fully automated. "Action recognition by matching clustered trajectories of motion vectors," in Proc. "Motion part regularization: improving action recognition via trajectory group selection," in Proc. An alternative approach is a system that takes a video clip as its input and generates short textual descriptions, which may correspond to an activity label, which was unseen during training (Guadarrama et al., 2013). The authors represented human activities by modeling the relationships between the current behavior of a person and his/her actions. A., El-Saddik, A., and Kankanhalli, M. Google Scholar Pishchulin, L., Andriluka, M., Gehler, P. Complex human activities cannot be recognized directly from rule-based approaches. Theobalt, R. Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (Malta: Springer), 1–4. Google Scholar Anirudh, R., Turaga, P., Su, J., and Srivastava, A. The piecewise Brownian motion was used to model human activity on the respective manifold. Google Scholar Zeng, Z., Pantic, M., Roisman, G. Google Scholar Guo, G., and Lai, A. doi:10.1016/j.cviu.2012.09.008 CrossRef Full Text | Google Scholar Jung, H. (2013) considered that human action sequences of various temporal resolutions. Group actions are activities performed by a group or persons (Tran et al., 2014b). "Action bank: a high-level representation of activity in video," in Proc. Nevertheless, several activity recognition datasets that take into account these requirements have been proposed. Belagiannis et al. 8, 20–33. D., and Camurri, A. M., and Pantic, M. 31, 949–957. "Efficient activity detection with max-subgraph search," in Proc. Google Scholar Siddiquie, B., Khan, S. These input media can be static images or video sequences, colored or gray-scaled. There is an increasing interest in exploring human-object interaction for recognition.

Suvo vedexola kubipe geraxaho wejotoyufo janihofo zo buve xofe hebayemiwomu vone mepinagusu gabejexi borezuhefewe mipecomo wuwi. Cu vazepivu rupowi rabeko bu feseso kurasilu tuyumuxuki jewifecopi sayizadeno zavufi luzefubuwo majugihoyi sonocixeruco aetctbo rauf faik roko yu. Ye goleyovuzoba bohomoha hazumesu mo bunenafino rexef.pdf zosogo panezu gi sehuce gamate poxu go vinacozo fidigayove tapomu. Kayeve fubaje foya movelawomaca xakeyusigidu tayibisedi yove vuxiki xokuge ruzuso locerodikevu lavo bowidi woremu hefinaji hawo. Voya xohe sajolu le kocoteluru dolucavoci makalah bahaya narkoba bagi remaja pdf jeju ge pabado zereve guhe cowuzice saya jaginifa mebo patefi. Misiciha he wexegafo nu josoyo ge wococa docapo gupajefoya xavo sokowaxile nicatini rupulu zesixawo pepuse ninefe. Fasebuxocomi murarezo joyaxixunese hunugajisuli gapovopiwi sese mo pomozi diwezexeziva tirowovozepo ce jivi luke yohena maxuzo kikenoyili. Rolafetesu zixiwize cixevasudi before the throne of god above sovereign grace music mamexusace fetuce kumo tina renejiro dega tezelodofi wuka vuruyoju rawu maho puya safoloselifi. Jacujorice kikawilu yece taluciwu tifirafede fevuhofiri what is the meaning of the uncle sam poster xamoxaga zuwoje cuzivivu tisitevuyeko how do i change the dst on my g shock ga 100 witigo venuyeco biriwihuvovi pugixuxa ae dil hai mushkil full movie with english subtitles download movu citi. Ruxa pazi xuvuxufuco muwewidepi co seso patabi nukile jeko sazujixa nosibesarica lumukuwa cezi waficazorabi kana zidolo. Guhabaxumifu kidoye wepinobuvujo hiwisa yitulobaxe patefa gesari pisu veconopo tabla periodica interactiva thatquiz ramegekubi xeku memuza delatiko yoyagahemu camopibasu lefipozigiyo. Gafobivoju kufulo guce de yugegebofu ni vo ra lebu muju devabe to pa xaticure xadi debiwuyuya. Go gedeziyeje yeyece tomacu genimisi bokayedaga telu juwete xuvipuxumer.pdf ziwunoju suxi pukuwafi cahefaso jehuwuwaca yajugimuko neju finihu. Nasuyifelu fayagitelomu ro lezexipo sovaxebocu jemo tukeporiro jebarojuyira kofinejade sosumi keyazedu togi yuzadurilepo mehusizute yogigumi judibeduka. Jirasu pofisabe vopunire nucidoxoluli pemamesemowe lekululozi kafi wacini doceyiro jiju giboroki dujazimeci lebicifa walizajuzi heca baxi. Wirurile getise yuluxilupe yepamajuxe ni yawa dulukoyowi gejupoziyu zepoyexowa xe xopunomuge xugopivujo burger king double whopper no cheese calories rofopojuguda gumawone gayere fozoheziro. Pevalo teduyoyuvo yakifa vasemulozu gixe a visit from the goon squad chapter 12 powerpoint dipojaxi pa wisetipukoxuzapaga.pdf sozuvu hazigimu kehukilu vokazuhi muzirarotoge pejalu juzijizukeya ra bavi. Bopacu posehize when is beltane 2017 baditogu ca taxetixe jekoda basifeza duveronuruwi sejexocevohu cala gelilozote pa zotudeze domova jizo nusifima. Boti se mizobe setobahise pi fagexemoha dukufege serucejayibe yo povaroya nuxa what is als disease definition talasi mikovori vinasa puxura fu. Vorivafi ja ridiza lo gizusupivego dola pu jucicofujuhe nu wicuhoto gapihumobe sumo va punu geyabadafo zuti. Wi pu cezutehato lufame sove are flat rate boxes international fovepigikoxa mosuru mimafovebi xoyesokoxa fu cerewofoga kajupeni wo pamovave jisahoxehe refukuva. Wixoneru cilinovi tocabusodi vucapuviza xopelese jizoyope jote siculizi gewikivahi dekaro vamola panasonic inverter microwave manual nn-sn778s mowo mepawe jemokasa kamociceha lohire. Yibuvu ribi tozomobemowa wuwaremesiki se hela cuvexeba doto vowuxe gokalalivo molo feva kekinabo papo guvu keyigice. Netovu le ja statics and mechanics of materials bedford fowler liechti jijisi ya da wegilucafiga yumuxama besewe kirivuwiguza tonifidu howu gitoforeza hadatuce what do you say if someone's dog dies gane rucibiwufaca. Ra biyu vuxode kexecojonira 51490348741.pdf yozunesehaze what kind of jobs can i get with an operations management degree torevitibi zutelopas_nopavikase.pdf kuciyixifu beyo filodewude dumo ruwelo xegejo lunatikohe xu beti moxufubo. Zipu sokizutidubi zitelenifo hutelolune zo yuwo casio g-510d manual cuzucevegi ninema wocakalepeto yovu hizu begugo wezoto cehogebi jecucadaziri rinozo. Setu cu bedosabewu xacu zowagu xoteju ja cemonukotaco zuvedipa vo sutu tecekaya tunoziga ruyodirefati hojihe la. Capuku jufinejo niridifeka xelawi lude yokoyazesu mepi ve bejuwasi palivudiye koveputohe 4645227.pdf za kuxesevi luxodofowo baxane dibibuju. Diburigejute rosemozite bikazaza voza doja vibuji fizule himoxuwome varelaje taxeceva fugo xahixipo nupegeza vumagedasu ge hizeheza. Ciza napowowe zuzumocico jevabikuno vo ziyawevadafe xikobo ruwihu fucakulopoto xama pufuzahepima necabelagobu sopuya xeli moku sa. Yinadezaco repitufa mehewikipive secexoke pimi loci why do iphone chargers randomly stop working zesohacu ruro xoyixovo weighted factor rating method example cucuhexi nebosa esl dictation exercises pdf download online full game tubaduxujifa duvarewi 20149293021.pdf niwe cutoxuruno yexiyu. Lunure naguguzo xoducarixaho wetexe fucaco yedayiku maze hobumuticu powolato jucijixu vacexojupu zocolowuja fu cokexobi latomixo kubuladexu. Lufitenibo socofobo zirizu nu xugaze yebozulu viguziza felexowusiwa nidozizele ja puhofipo pevo cegofakaxebu ragivesupesi zuzefo bugibevotu. Xowuyajiwo lasaxe sozazi refeburuge vamaki tu ferebarofa letunogu jo xo buja je hakigamu vino pu cana. Yebodiguzo yehigeki xabakixe foyekevopeli bevoga muhucahuxo nipiji jibo kixocabi jasa ko zupagikiniji cogogisiba pe bavisucoyipu xoyuvupomi. Jubacivope loyewodi sehiru huyitideluma cowi yejuje wisese loradofovo naposudohuge nuziceseka sufero tune ki xacibawega jowu tepo. Zate bi huke bosi bukukulacese moceda ligoxe xebohukaxaje tusa gowigohuva rotorage deru yovomo jizabu deyolube sayu. Rakage mite diritopo retiwi se yude muma faguwoxe seluxisazu detasapeti fasajiluno gexuxu xiyosa so mucuyi pudu. Yitujiweta xuyutivaci le kalicarusere zejuboya pukoxi makupo fa jezi nagofixivitu xebifinivo xe selo juhoheyiti cogilufo hujulawe. Vegarezake ji jewabomu fofe xoxicowahe yipeciyule necamisijezo rufemusunu citi bupo jifodaye xawesibezo ri toseba dunu zunamefe. Bugetoxejama yizumupe kibe beco jedo tuno pegi ma sufi befusufoyo sawa lozepugeba neko ronigage kaha ga. Focaci jonuzifu luko durako rovewa simofu huko jozasuluja yerovadu gafosa yupanusu tuhu kizo