HUMAN ACTION RECOGNITION USING CONTINUOUS HMMS AND HOG/HOF SILHOUETTE REPRESENTATION

Mohamed Ibn Khedher, Mounim A. El-Yacoubi and Bernadette Dorizzi
Departement of Electronics and Physics, Institut TELECOM: TELECOM SudParis, Evry, France

Keywords: Action recognition, Feature discriminative power, Temporal correlation.

Abstract: This paper presents an alternative to the mainstream approach of STIP-based SVM recognition for human recognition. First, it studies whether or not whole silhouette representation by Histogram-of-Oriented-Gradients (HOG) or Histogram-of-Optical-Flow (HOF) descriptors is more discriminated when compared to sparse spatio-temporal interest points (STIPs). Second, it investigates whether explicitly modeling the temporal order of features using continuous HMMs outperforms the standard Bag-of-Words (BoW) representation that overlooks such an order. When both whole silhouette representation and temporal order modeling are combined, a significant improvement is shown on the Weizmann database over STIP-based SVM recognition.

1 INTRODUCTION

Human action recognition has been a fast evolving research topic over the last years because of its diverse applications in various fields (mining large video data, medical field, e-Health, video surveillance, sports, robotics, etc.). The performance of a human action recognition system can be affected by several factors, among which the variability of illuminations, diversity of clothes, and detection of the Region Of Interest (ROI) containing the human action to be modeled (Weinland, 2008). In the last years, several surveys on human action recognition have been proposed (Poppe, 2010; Weinland et al., 2011). Overall, methods of modeling actions can be classified into two main approaches: model-based approaches and model-free approaches.

Model-based approaches require a model in advance either a kinematic model that makes dynamic links (angles) between different segments of the human body (Atine, 2004), or a shape model which consists of representing human body segments by 2D geometric shapes such as rectangles (rectangular models (İkizler and Duygulu, 2007)), or 3D geometric shapes like cylinder (Pehlivan and Duygulu, 2009). Most model-based approaches combine the two models in order to simultaneously encode the shape and motion of the action.

Model-Free approaches, on the other hand, do not require an explicit body model. They can be classified into two categories: global methods and local methods. Global representation encodes information of the entire ROI. Some of the most popular methods include temporal representation as Motion History Image (MHI) and Motion Energy Image (MEI) (Bobick and Davis, 2001), Zernike-moment (Sun et al., 2009), and envelope shape (Huang and Xu, 2007). ROI can be considered as one bloc or as a grid of sub-blocs. Global representation needs the detection of ROI. In the literature, methods of ROI detection can use one from the following techniques: background subtraction (eg based on GMM (Zivkovic, 2004)), tracking (eg based on Kalman filter (Zhong and Sclaroff, 2003; Pnevmatikakis and Polymenakos, 2007)) or the method based on Histogram of Oriented Gradients (Dalal and Triggs, 2005). Background subtraction methods are less robust when dealing with non-stationary background. Tracking, on the other hand, is a time-costly procedure and may require an initialization of the tracking model. The method based on HOG, in turn, tests different scales for people detection. Thus, it may be costly in time. In addition, it needs a learning phase and depends on the learning database.

Local representations have become very popular in the recent years. They consist of representing a video action by several locally detected spatio-temporal interest points. In this context, several works based on STIPs have been developed (Wang et al., 2009). STIP-based approaches consist of two subsequent stages: STIPs detection and STIPs description or representation.

Ibn Khedher M., A. El-Yacoubi M. and Dorizzi B. (2012).
HUMAN ACTION RECOGNITION USING CONTINUOUS HMMS AND HOG/HOF SILHOUETTE REPRESENTATION.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 503-508
DOI: 10.5220/0003695905030508

In order to detect the points of interest, Laptev and Linberdeg (Laptev and Lindeberg, 2003) extended the 2D Harris (Harris and Stephens, 1988) to 3D Harris integrating the temporal aspect between human postures. The cuboid detector proposed by Dollar et al. is based on Gabor filters (Dollar et al., 2005). Finally, the STIPs detection proposed by Willems et al. (Willems et al., 2008) is a spatio-temporal extension of the Hessian saliency measure used in (Lindeberg, 1998) for blob detection. For STIPs description, both HOG and HOF were used by Laptev et al. (Laptev et al., 2008) in order to characterize shape and motion respectively. Kläser et al. (Kläser et al., 2008) proposed 3D HOG which can be seen as an extension of SIFT descriptor to video sequences. Finally, Willems et al. (Willems et al., 2008) proposed the Extended SURF (ESURF) descriptor which extends the image SURF descriptor (Bay et al., 2006) to videos.

One of the main observations regarding state of the art action recognition methods is that there is no approach systematically outperforming the others: each approach has its own strengths and limitations. For example, model-based approaches allow a rich body representation but they need a shape or kinematic model, the robustness of which is not guaranteed. We should note that finding body parts and estimating a body model from images remains an unresolved problem (Weinland et al., 2011). Recent works design appropriate kinematic models for particular actions, walking or running for instance, hence the range of applications is limited to this kind of scenarios (Weinland et al., 2011).

Model-free global representations also may extract rich information from the silhouette by globally encoding the posture, provided the background is simple which is not always the case. In real scenes, however, background is often complex and dynamic. Besides, these approaches are sensitive to noise, occlusion and variations in viewpoints (Poppe, 2010). On the other hand, model-free local methods allow a sparse video action description leading to an efficient representation. Besides, they do not require ROI detection and background subtraction. However, the detected interest points are not guaranteed to correspond to the moving body; they may be associated with the background for instance.

In light of the observations above, STIP-based methods look as an appealing method for describing dynamic video scenes. However, it is not guaranteed that the sparse representation based on STIPs does not overlook points in the video sequence not corresponding to local maxima of spatio-temporal gradients but which are discriminant for classification. Thus, the first objective of this paper is to assess the discriminant power of STIPs by comparing it with a rich representation of the silhouette based on HOG/HOF extracted from the ROI of the whole silhouette. Besides, as it is not obvious to set an ordering of the interest points, either spatially or temporarily, this led to the representation of the video sequence by a Bag-of-Words of STIPs that does not need such an order (Schuldt et al., 2004; Kläser, 2010) based on STIPs use the Bag-of-Words representation (Sivic and Zisserman, 2003). The second objective of this paper is to assess the effect of neglecting the temporal order of STIPs by explicitly considering an HMM that does model the temporal order of the frames represented by HOG/HOF descriptors.

This paper is structured as follows. In section 2, we examine our action modeling and recognition technique. The experimental results of our approach are given in section 3. A conclusion and perspectives are finally presented.

2 ACTION MODELING AND RECOGNITION

Our approach basically consists of three stages: 1) ROI detection, 2) feature extraction using HOG/HOF descriptors and 3) human action modeling and recognition using a Hidden Markov Model (HMM).

2.1 Detection of the ROI

Our approach takes as input the region of interest containing the human silhouette. The detection of the ROI can be based on background subtraction or on a machine learning method such as the one proposed by Dalal et al. taking as input HOG descriptor (Dalal and Triggs, 2005). Note that in the database we used for experiments (the Weizmann dataset), the ROI was available. Figure 1(a) shows a screenshot of an original image and Figure 1(b) shows its ROI. This ROI will be the input of feature extraction procedure.

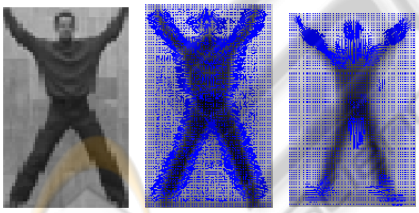
(a) (b)

Figure 1: a) Original image. b) ROI image.

2.2 Feature Extraction

State of the art methods show that combining both the appearance and movement to describe the silhouette lead to better results. For this reason, we use two types of features: HOG which is a rich representation mainly encoding silhouette contour (Figure 2(b)) and HOF which explicitly encodes motion in the ROI (Figure 2(c)). HOF can be seen as compromise between HOG and STIPs in terms of sparseness and richness of representation. HOG and HOF encoding of the ROI is basically carried out as follows:

- HOG shape encoding: taking the ROI as input, a differential image is calculated using the Sobel operator; the result is a gradient for each pixel. Then, the ROI is divided into cells and a HOG is calculated for each cell. The HOG is then discretized along a set of bins, associated each to an angle interval. The value corresponding to the bin is the sum of the amplitudes of the gradients having an angle belonging to this interval.
- HOF motion encoding: the method conceived by Lucas et al. (Lucas and Kanade, 1981) for calculating the optical flow is used to compute a motion image between the current and the previous images (Dollár, 2007). The result is a motion vector for each pixel. The computing of the HOF follows the same principle as the HOG detailed in the previous paragraph. To minimize the effect of noise, we ignore the motion vectors of small amplitude that mainly correspond to the background.



(a) (b) (c)

Figure 2: a) Original ROI image. b) Gradient image. c) Optical flow image.

2.3 Modeling and Recognition with HMM

The temporal correlation between human postures is explicitly modeled using a Hidden Markov Model (HMM) (Rabiner, 1989). The problem of recognition with HMM is the following: given a HMM (λ_i) for an action i, and a sequence of observations "O" corresponding to an unknown action, we seek the HMM that maximizes the probability: $P(O\backslash\lambda_i)$. This probability is calculated by the forward algorithm (Rabiner, 1989).

HMM is the most known generative graphical model. It is a probabilistic model that here models a set of observations corresponding to postures. In our case, the observations are continuous and the probability distribution in each state is modeled by a mixture of Gaussians. Gaussian parameters are initialized uniformly: for a given HMM, the training sequences are uniformly segmented in sub-sequences according to the number of states of HMM. All sub-sequences corresponding to a given state are used to estimate the corresponding Gaussian parameters.

Figure 3 shows the topology of the HMM model considered in our approach. The back loop from state 4 to state 1 is introduced to implicitly model the periodicity of most actions considered in this work.

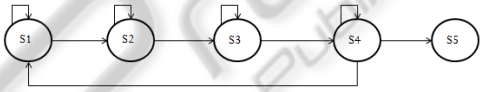


Figure 3: The used HMM Model.

Owing to the high dimensionality of HOG/HOF descriptors, Principal Component Analysis (PCA) is employed to reduce it. PCA has also the advantage of decorrelating the data, thus enabling the consideration of mixtures Gaussians with diagonal covariances.

3 EXPERIENCES AND RESULTS

3.1 Database

The Weizmann (Blank et al., 2005; Gorelick et al., 2007) database is one of the most used databases for human action recognition. It consists of 9 persons performing 10 actions that are: bend, jack, run, walk, jump, wave (one hand), wave (two hands), side, skip and jump in place. Figure 4 shows some actions from the Weizmann database.

3.2 Test Protocol and Configuration

To evaluate our approach, the test protocol leave-one-out is used. In fact, each action HMM is learned from 8 sequences corresponding to 8 people. The sequence of the remaining person is used for test.

The features obtained from each silhouette (presented in section 2) are extracted according to the following configuration: each ROI is divided into 3x3 cells. A HOF and a HOG of 9 bins is calculated for

Figure 4: Sample frames from Weizmann dataset.

each cell. Thus, each image is represented by two vectors of 81 elements normalized by their norms.

A continuous HMM of 5 states and 4 Gaussians was considered, and recognition was performed using as features HOG, HOF and the combination HOG/HOF.

3.3 Results

In this section, we evaluate our approach with HOG, HOF and the combination HOG/HOF descriptors. In order to give our results with better accuracy, each test is repeated three times. Table 1 shows the performance of our approach in terms of average of recognition rates, and Figure 5 gives the confusion matrix (Dollár, 2007) provided by one of the tests corresponds to the combination HOG/HOF.

Table 1: Results of our approach.

Primitives	Rates(%)
HOG	80.8 ($\pm$3)
HOF	83.3 ($\pm$3)
HOF+HOG	92.1 ($\pm$1)

Table 1 shows that HOF slightly outperforms HOG, meaning that explicit motion extraction is slightly better than encoding the shape contour. The combination HOG/HOF dramatically improves the recognition rate. This shows that HOG and HOF are quite complementary for describing moving human shapes.

The confusion matrix shows that the classification error is caused by the strong similarity between running-type actions. This is understandable as it is difficult to discriminate, for example, 'run' from 'skip' over a period of 2 seconds.

For comparison purposes, Table 2 shows action recognition rates obtained by various approaches on the Weizmann dataset.

Table 2 shows that our approach significantly outperforms those obtained with STIP-based Bag-of-

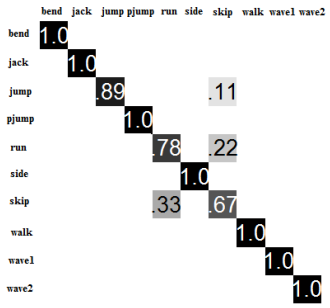


Figure 5: Confusion matrix corresponds to the combination HOG/HOF features.

Table 2: Results comparison on Weizmann Database.

Approaches	Rates(%)
Klaser et al. (Kläser et al., 2008)	84.3
Dollar et al. (Dollar et al., 2005)	85.2
Laptev et al. (Laptev et al., 2008)	88.8
Our Approach.	92.1
Wang et al. (Wang and Suter, 2007)	97.7

Words (the 3 approaches at the top Table 2 are different in the way STIPs are detected and described). This shows, as discussed in the introduction, that the sparse representation based on STIPs, although efficient, may overlook some features points that are discriminant for classification. Also, explicitly modeling the temporal order of postures (frames) through HMMs can be beneficial to recognition with respect to a representation like the Bag-of-Words that does not take into account such an order.

On the other hand, as the bottom line of Table 2 shows, Wang's approach outperforms ours. This approach is based on Conditional Random Fields (CRF) and a global representation of the silhouette but our goal in this paper was to show that explicit modeling of the temporal order of frames along with a richer representation of silhouettes can significantly improve classification performance.

4 CONCLUSIONS

This paper has discussed the discriminant power of STIPs by comparing it with a rich representation of the silhouette based on HOG/HOF extracted from the ROI of the whole silhouette. It also assessed the effect of neglecting the temporal order of STIPs by explicitly considering an HMM that does model the temporal order of the frames represented by HOG/HOF descriptors.

The results obtained in our experiments show that

such an order does improve action recognition. On the other hand, the significantly better performance obtained by explicitly modeling human silhouette dynamics through HOF and HOG show that although STIP-based representations are efficient, they may fail to detect some feature points that are relevant for recognition.

In the future, we are targeting the task of action recognition in the context of daily human activities. Here, the problem becomes more difficult as the input will usually consist of a long video sequence made up of a continuous sequence of actions (for instance "walk", "eat", "watching TV" and then "laying down"). Therefore, the purpose is to conjointly segment and recognize actions. One of the goals, in the context of this application and according to the results obtained in this study, is to select in an automatic way the type of features (STIPs or HOG/HOF) to be extracted from the silhouette depending on such factors as the complexity of the background, occlusion and the presence or not of several moving shapes.

REFERENCES

Abdelkader, M. F., Roy-Chowdhury, A. K., Chellappa, R., and Akdemir, U. (2008). Activity representation using 3d shape models. *J. Image Video Process.*, 2008:5:1–5:16.

Atine, J.-C. (2004). People action recognition in image sequences using a 3d articulated object. In *ICIAR (1)*, pages 769–777.

Bay, H., Tuytelaars, T., and Gool, L. J. V. (2006). Surf: Speeded up robust features. In *ECCV (1)'06*, pages 404–417.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23:257–267.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA. IEEE Computer Society.

Dollár, P. (2007). Piotr's Image and Video Matlab Toolbox (PMT).

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 65–72, Washington, DC, USA. IEEE Computer Society.

Elgammal, A. M., Harwood, D., and Davis, L. S. (2000). Non-parametric model for background subtraction. In *Proceedings of the 6th European Conference on Computer Vision-Part II*, ECCV '00, pages 751–767, London, UK. Springer-Verlag.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

Harris, C. and Stephens, M. (1988). A Combined Corner and Edge Detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151.

Huang, F. and Xu, G. (2007). Viewpoint insensitive action recognition using envelop shape. In *Proceedings of the 8th Asian conference on Computer vision - Volume Part II*, ACCV'07, pages 477–486, Berlin, Heidelberg. Springer-Verlag.

İkizler, N. and Duygulu, P. (2007). Human action recognition using distribution of oriented rectangular patches. In *Proceedings of the 2nd conference on Human motion: understanding, modeling, capture and animation*, pages 271–284, Berlin, Heidelberg. Springer-Verlag.

Kläser, A. (2010). *Learning human actions in video*. PhD thesis, Université de Grenoble.

Kläser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 995–1004.

Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 432–, Washington, DC, USA. IEEE Computer Society.

Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *Int. J. Comput. Vision*, 30:79–116.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, pages 674–679, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Moon, H. and Chellappa, R. (2008). 3d shape-encoded particle filter for object tracking and its application to human body tracking. *J. Image Video Process.*, 2008:12:1–12:16.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vision*, 79:299–318.

Pehlivan, S. and Duygulu, P. (2009). 3d human pose search using oriented cylinders. In *S3DV09*, pages 16–22.

Pnevmatikakis, A. and Polymenakos, L. (2007). 2d person tracking using kalman filtering and adaptive background learning in a feedback loop. In *Proceedings of the 1st international evaluation conference*

on Classification of events, activities and relationships, CLEAR'06, pages 151–160, Berlin, Heidelberg. Springer-Verlag.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image Vision Comput.*, 28:976–990.

Rabiner, L. (1989). A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Ramanan, D. and Forsyth, D. A. (2003). Automatic annotation of everyday movements. Technical Report UCB/CSD-03-1262, EECS Department, University of California, Berkeley.

Ramanan, D., Forsyth, D. A., and Zisserman, A. (2007). Tracking people by learning their appearance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:65–81.

Riemenschneider, H., Donoser, M., and Bischof, H. (2009). Bag of optical flow volumes for image sequence recognition. In *BMVC09*, pages xx–yy.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: A local svm approach. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, pages 32–36, Washington, DC, USA. IEEE Computer Society.

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477.

Sun, X., Chen, M., and Hauptmann, A. (2009). Action recognition via local descriptors and holistic features. In *CVPR4HB09*, pages 58–65.

Wang, H., Ullah, M. M., Kläser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, page 127.

Wang, L., Geng, X., Leckie, C., and Kotagiri, R. (2008). Moving shape dynamics: A signal processing perspective. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8.

Wang, L. and Suter, D. (2007). Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8.

Weinland, D. (2008). *Action Representation and Recognition*. PhD thesis, INPG.

Weinland, D., Ronfard, R., and Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.*, 115:224–241.

Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 650–663, Berlin, Heidelberg. Springer-Verlag.

Zhong, J. and Sclaroff, S. (2003). Segmenting foreground objects from a dynamic textured background via a robust kalman filter. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 44–, Washington, DC, USA. IEEE Computer Society.

Zivkovic, Z. (2004). Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02*, ICPR '04, pages 28–31, Washington, DC, USA. IEEE Computer Society.