ACTION RECOGNITION USING SPATIO-TEMPORAL DIFFERENTIAL MOTION

Gaurav Kumar Yadav, Amit Sethi

Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati
Guwahati, India

ABSTRACT

This paper presents human action recognition using spatio-temporal differential motion maps. The concept of differential motion in space and time helps in overcoming several challenges in action recognition such as camera motion and multiple actions in the same scene. Spatially differential motion in a frame is represented using divergence of optical flow. Divergence map of each frame in a video is projected onto three orthogonal Cartesian planes. A map of spatio-temporal differential motion is formed by accumulation of the absolute differences between projected maps of pairs of consecutive frames through an entire video sequence for each projection. A feature vector is formed from these three spatio-temporal maps of differential motion which represents the action performed in the video. Classification of action was done by using l_2-regularized collaborative representation with a distance-weighted Tikhonov matrix. We tested on two popular datasets, KTH and UCF11, and got better performance than state-of-the-art methods. A comparison of differential motion and optical flow (any motion with respect to the camera) was also done to show that differential motion gives better feature representation than simply using optical flow.

Index Terms— Differential motion maps, action recognition, optical flow, Divergence

1. INTRODUCTION

Human action recognition is a challenging problem in the field of computer vision. In daily life video data is increasing day-by-day. Videos are captured from many devices such as mobile phones and cameras. Video captured with two different devices shows lots of variations. Because it is very tedious to annotate videos that are easily captured, automated classification and annotation is very much required while dealing with large video databases. That is why human action recognition has become popular in video classification and video retrieval problems. There has been a lot of research in the past years on this problem but a real-time solution has been not proposed yet. Most of the state-of-the-art methods can only achieve good performance on videos captured in constrained environments.

There are various challenges in human action recognition such as variations in environments, viewpoints and actor movement. Variations in environments are caused by moving background, occlusion and addition of noise while capturing the video. Environment and recording settings also causes various types of noise in different lighting conditions. Videos of the same action class taken from different viewpoints have high intra-class variation. Such intra-class variation is comparable to inter-class variation for classes of similar actions such as walking, jogging and running, which differ mostly in speed and stride length. To achieve a real-time performance a system must be robust to changes in illumination, viewpoint, actors, and environment.

Human action recognition has become a popular research topic mainly because of its application in video classification and surveillance. Action recognition methods can be divided into two categories: silhouettes-based and motion-based.

Silhouettes-based approaches extract the silhouettes or skeletons to recognize the action. In [1], the authors proposed a method based on contour points of the silhouettes to represent different poses. Pose learning was done using k-means clustering and Euclidean distance. Wu et al. [2] exploited the correlation between poses and bag-of-words model for feature extraction. A viewpoint-independent silhouette-based human action recognition was proposed by Orrite et al. [3]. Each action template was projected onto a new subspace by means of the Kohonen self-organizing feature map and action recognition is accomplished by a maximum likelihood (ML) classifier. In all the shape-based methods contour points need to be detected to form the skeleton or to capture the pose. Due to fractured silhouettes and overlapping body parts exact shape extraction is difficult. Also skeleton information is not available in most of the cases.

Motion-based approaches are much simpler and easy to compute. In [4, 5] an interest point detector and SVM was used for action recognition. Dollar et al. [6] proposed an efficient approach for detecting spatio-temporal interest points using a temporal Gabor filter and a spatial Gaussian filter. In [7], wang et al. proposed a trajectory-based feature for action recognition. Feature extraction was done by forming group of trajectories and computing histogram of oriented-gradient (HOG), histogram of flow (HOF) and motion-boundary his-

togram (MBH). In case of space-time approach, generally features were extracted from a volume around spatio-temporal interest points and bag-of-words [8] model is used to represent the activity. Weinland et al. [9] proposed motion history volumes for feature extraction and Mahanolobis distance was used for classification.

Recently, there has been a lot of advancements in the field of deep learning. Deep learning methods give good performance for the large number of training samples. In this paper, we focus on small datasets where deep learning methods fails to achieve good performance.

We have developed a method for human action recognition based on the observation that information about an action is contained in differential motion between objects in space and time. Our contributions are as follows:

Differential motion: Proposed a method for computation of differential motion. It captures the motion of the moving objects with respect to a potentially non-stationary background very effectively.

Differential motion maps: Proposed a feature representation of video based on the differential motion maps for classification of actions.

Differential motion vs. optical flow: We experimented with both differential motion and optical flow, that is, any motion with respect to the camera. Feature representation was computed for both and it was found that differential motion gives better performance than optical flow.

2. PROPOSED METHOD

Our main idea for capturing unique features of each action class is to capture motion between an actor and the background, as well as changes in that motion from frame to frame. Optical flow is not enough to distinguish between different action classes, especially when the camera is moving because it gets overwhelmed by the motion of the background with respect to the camera. On the other hand, spatio-temporal points of high acceleration represent change in the direction of motion, or onset or end of an atomic action. Such points are vital in defining an action. For example, in clapping, the moment when the hands start moving towards each other, or when they suddenly stop after coming together are iconic. Therefore, the proposed algorithm is based on the computation of differential motion. Differential motion maps capture structure and shape information.

The proposed method has two major parts: feature extraction and classification.

2.1. Feature extraction

To compute motion information we use optical flow based on Lucas-Kanade method [10]. For each frame in the video, the flow is computed. Optical flow (velocity) vector $\overrightarrow{V}$ of a point in a frame is expressed as a linear combination of unit vectors in $\overrightarrow{i}$ and $\overrightarrow{j}$ in x and y directions respectively as $V_1\overrightarrow{i} + V_2\overrightarrow{j}$. Divergence is a differential operation on the vector field defined by flow map as follows:

$$R(x,y,t) = \nabla.\overrightarrow{V}(x,y,t) = \frac{\partial V_1(x,y,t)}{\partial x} + \frac{\partial V_2(x,y,t)}{\partial y} \tag{1}$$

The magnitude of divergence of optical flow is usually high along the edges of a moving object when those edges are moving perpendicular to the tangent of the edge relative to their background, as shown in Fig. 1(b). This is because there is unidirectional motion on one side of the point on the edge, which gives a net non-zero contribution to the line integral used for computing divergence around the point. Then, a point-wise absolute difference of divergence of pairs of consecutive frames as shown in Equation 2 to capture the change in motion, where, without loss of generality, we assume that the frame time stamps are represented by integers.

$$D_f(x,y,t) = |R(x,y,t+1) - R(x,y,t)| \tag{2}$$

To compress the information, we project the divergence map for each frame onto the three orthogonal Cartesian planes defined by (x,y,t) coordinates as proposed in [11]. This absolute difference of divergence projected to the three planes is accumulated for each pair of consecutive frames for the entire video. Thus, the entire video is used to generate three 2D projected maps corresponding to front, side and top views as shown in Fig. 1. These differential motion maps are analogous to, yet different from, depth motion maps [11]. For the front view in which the dimension t is eliminated, this operation is defined as:

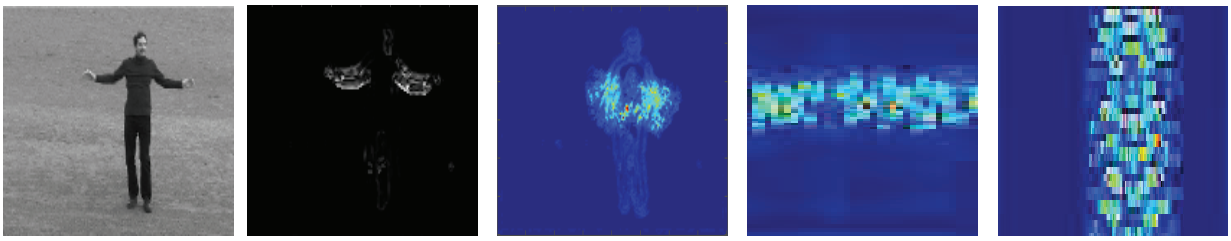
$$DMM_{front}(x,y) = \sum_{t=1}^{T-1} D_f(x,y,t) \tag{3}$$

Similarly for Side and top view:

$$DMM_{side}(y,t) = \sum_{x=1}^{m} D_f(x,y,t) \tag{4}$$

$$DMM_{top}(t,x) = \sum_{y=1}^{n} D_f(x,y,t) \tag{5}$$

To deal with different sizes of motion maps for different actions, bi-cubic interpolation was used to resize all projection views to a fixed size. All values of the maps were normalize between 0 and 1 to avoid large values dominating the feature set, and to normalize for video length N. The size of DMMs for front side and top views were $m_f \times n_f$, $m_s \times n_s$, $m_t \times n_t$ respectively. For an action video sequence, a feature vector of size $(m_f \times n_f + m_s \times n_s + m_t \times n_t) \times 1$ was formed by concatenating the vectorized differential motion maps of three views.



(a) (b) (c) (d) (e)

Fig. 1. (a) Example frames for hand-clapping video sequence, (b) divergence magnitude of optical-flow (c) Front view ($xy-$plane) differential motion map, (d) Side view ($yt-$plane) differential motion map, (e) Top view ($xt-$plane) differential motion map.

From differential motion maps in Fig. 1 we can easily see that the movements of hands have been captured for hand-clapping video sequence in case of front, side and top views. In Fig. 1(c), horizontal axis denotes the width and vertical axis denotes the height of the video sequence and high values at the center shows the hand-clapping motion from front-view. Similarly in Fig. 1(d), horizontal axis denotes the time dimension and vertical axis denotes the height of the video sequence and high values along the horizontal direction shows the position of hand motion from side-view. In Fig. 1(e), horizontal axis denotes the width of the video sequence and vertical axis denotes the time dimension and high values along the vertical direction shows the position of hand motion from the top-view.

2.2. Classification

As evident from our feature design, the feature is a linear additive mixture of individual features associated with various atomic actions in the videos due Equations 2–5. For classification l_2-regularized collaborative classifier [11] (LRCC). LRCC is based on sparse representation coding. Consider a dataset with C classes of training samples arranged column-wise $\mathbf{A} = [\mathbf{A_1}, \mathbf{A_2}, ..., \mathbf{A_c}] \epsilon R^{d \times n}$, where $\mathbf{A_j} (j = 1, ..., C)$ is the subset of the training samples associated with class j, d is the dimension of training samples and n is the total number of training samples from all the classes. A test sample $\mathbf{g} \epsilon R^d$ can be represented as a sparse linear combination of the training samples, which can be formulated as $\mathbf{g} = \mathbf{A}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_C]$ is an $n \times 1$ vector of coefficients corresponding to all the training samples and $\boldsymbol{\alpha}_j (j = 1, ..., C)$ denotes the subset of the coefficients associated with the training samples from the j^{th} class, i.e. $\mathbf{A}_j$. The solution for $\boldsymbol{\alpha}$ can be found by solving equation 3 using l_2 regularization as follows:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \|\mathbf{g} - \mathbf{A}\boldsymbol{\alpha}\|_2^2 + \lambda \|\mathbf{L}\boldsymbol{\alpha}\|_2^2 \right\} \tag{6}$$

where λ is regularization parameter and $\mathbf{L}$ is the Tikhonov regularization matrix. The class label of $\mathbf{g}$ can be obtained from equation 4 as proposed in [12] as follows:

$$class(\mathbf{g}) = \arg\min_j (\mathbf{e}_j) \tag{7}$$

where $\mathbf{e}_j = \left\| \mathbf{g} - \mathbf{A}_j \hat{\boldsymbol{\alpha}}_j \right\|_2$.

3. EXPERIMENTS AND RESULTS

3.1. Experimental setup

Two popular datasets were used for the experiments. KTH is simple action database which consist of 6 actions including boxing, hand-clapping, hand-waving, jogging, running and walking [4]. UCF11 is a complex action database which consists of 11 actions including basketball shooting, biking, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog [13]. For KTH dataset we followed the same procedure as mentioned in [4]. We divided the dataset into training and testing sets based on the subjects. For UCF11, It has been suggested to use entire groups instead of splitting their constituent clips into training or testing sets (specifically, leave-one-group-out cross validation (LOOCV)) to test robustness of action recognition techniques to unseen scene variations [13]. Classification of actions was done using LRCC. The accuracy was averaged over several random selections of training and testing data. For UCF11 dataset we have followed the LOOCV approach, which leads to 25 cross-validation results for each action class, which were averaged.

Principal component analysis (PCA) was applied to reduce the dimensionality of the features. The PCA transform matrix was obtained using the training feature and then applied to the test features. The largest eigenvalues that explained 85% of the variance were kept. This reduced the dimension from approximately 17,000 to 400. Dimensionality reduction improved the computational efficiency without affecting the performance. We have considered the dimension after which accuracy becomes constant. Regularization parameter λ was tuned by grid search using cross-validation on training datatsets.

Table 1. Recognition accuracy for KTH and UCF11 datasets.

Method	KTH	UCF11
Proposed method	96.98%	90.24%
Yadav et al. [14]	98.2%	91.3%
Kovashika et al. [15]	94.53%	90.45%
Gilbert et al. [16]	94.50%	–
Wang et al. [7]	94.20%	84.20%
Laptev et al. [17]	91.80%	–
Shuiwang et al. (CNN) [18]	90.2%	–
Mahdyar et al. (CNN) [19]	–	89.5%
kizler-Cinbis et al. [20]	–	75.21%
Liu et al. [13]	–	71.20%

Table 2. Comparison of optical flow and differential motion for KTH and UCF11 datasets.

Dataset	Optical flow	Differential motion
KTH	65.00%	96.98%
UCF11	44.10%	90.24%

3.2. Recognition results

Comparison with the state-of-the-art: We compare our results with the state-of-the-art methods, as shown in Table 1. The proposed method performed better than the state-of-the-art methods for KTH as well as UCF11 datasets. LRCC performs better because LRCC models a video as a linear combination of several prototype feature vectors from the training data. This allows it to decompose a videos feature vector into its components arising from both in-class and out-of-class actions in the same video. The coefficient of in-class component helps it classify better for complex videos with out-of-class actions such as those in UCF11. In line with several contemporary techniques, scale and view invariances are ensured by the diversity of training patterns but not explicitly per se. Empirically, we have demonstrated using UCF11 that our method can deal with action in large frames that may even include additional out-of-class actions in the background. Our method performs outperformed CNN-based methods [18, 19] because CNN requires large number of samples for training also training time is more in case of CNN. Here, we focus on small dataset where CNN does not perform well.

Impact of differential motion: If we directly used optical flow, the results were not as good as using divergence of optical flow, keeping the rest of the pipeline the same, as shown in Table 2. This shows that differential motion provides better features than absolute motion. Therefore, by using differential motion we suppress small motion components which are noisy and keep only large motion components for classification.

Class separation: To show that proposed features are discriminative we computed the mean-square distance between inter-intra class as shown in Fig. 2. For KTH, It is clearly visible from the figure that the overlap between inter-intra class features was small which leads to better performance. As UCF11 is complex dataset, features are less discriminative for some actions such as diving, soccer juggling, trampoline jumping and walking, which affected the performance of classifiers. Also, intra-class feature distance was less compared to inter-class distance which justifies the class separation.

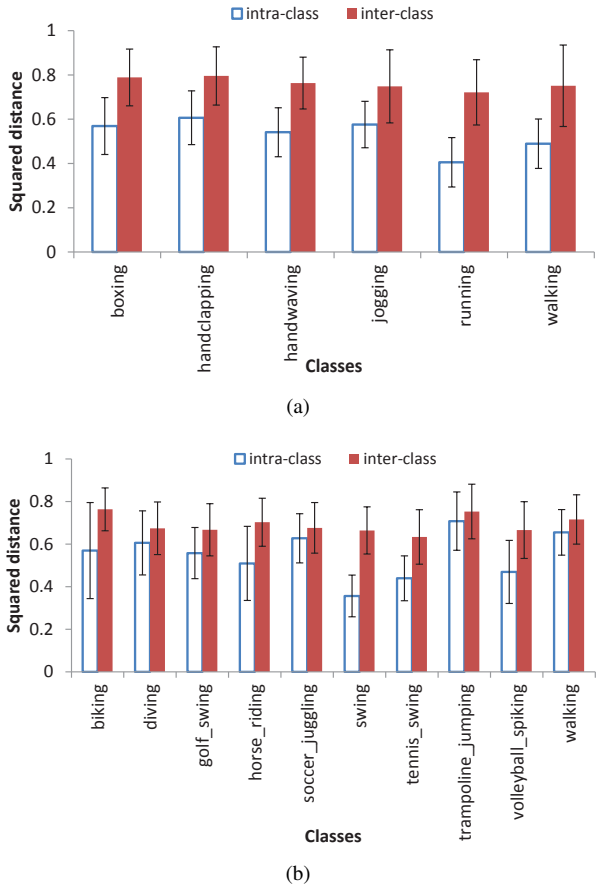


Fig. 2. Inter- and intra-class mean squared distances (and their variances) for the proposed video representation for (a) KTH and (b) UCF11 datasets.

4. CONCLUSIONS AND DISCUSSION

We proposed a feature representation based on differential motion map for action recognition. Differential motion captures the motion information very effectively and shows better performance compared to state-of-the-art methods. Differential motion maps capture the action structure as well as motion. A comparison of differential motion and optical flow based feature was also done to show that differential motion gives better feature representation. The proposed method was tested on two popular datasets KTH and UCF11. Our future goal is to test it on larger datasets.

5. REFERENCES

[1] Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[2] Di Wu and Ling Shao, "Silhouette analysis-based action recognition via exploiting human poses," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 236–243, 2013.

[3] Carlos Orrite, Francisco Martínez, Elías Herrero, Hossein Ragheb, and Sergio Velastin, "Independent viewpoint silhouette-based human action modelling and recognition," in *The 1st International Workshop on Machine Learning for Vision-based Motion Analysis-MLVMA'08*, 2008.

[4] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. IEEE, 2004, vol. 3, pp. 32–36.

[5] Gaurav Kumar Yadav and Amit Sethi, "A flow-based interest point detector for action recognition in videos," in *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*. ACM, 2014, p. 41.

[6] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72.

[7] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3169–3176.

[8] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[9] Daniel Weinland, Rémi Ronfard, and Edmond Boyer, "Motion history volumes for free viewpoint action recognition," in *Workshop on modeling People and Human Interaction (PHI'05)*, 2005.

[10] Bruce D Lucas, Takeo Kanade, et al., "An iterative image registration technique with an application to stereo vision.," in *IJCAI*, 1981, vol. 81, pp. 674–679.

[11] Chen Chen, Kui Liu, and Nasser Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *Journal of real-time image processing*, pp. 1–9, 2013.

[12] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[13] Jingen Liu, Jiebo Luo, and Mubarak Shah, "Recognizing realistic actions from videos in the wild," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1996–2003.

[14] Gaurav Kumar Yadav, Prakhar Shukla, and Amit Sethfi, "Action recognition using interest points capturing differential motion information," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1881–1885.

[15] Adriana Kovashka and Kristen Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2046–2053.

[16] Andrew Gilbert, John Illingworth, and Richard Bowden, "Action recognition using mined hierarchical compound features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 883–897, 2011.

[17] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[19] Mahdyar Ravanbakhsh, Hossein Mousavi, Mohammad Rastegari, Vittorio Murino, and Larry S Davis, "Action recognition with image based cnn features," *arXiv preprint arXiv:1512.03980*, 2015.

[20] Nazli Ikizler-Cinbis and Stan Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *European conference on computer vision*. Springer, 2010, pp. 494–507.