Human Action Recognition Using LBP-TOP as Sparse Spatio-Temporal Feature Descriptor

Riccardo Mattivi and Ling Shao

Philips Research, Eindhoven, The Netherlands
{riccardo.mattivi,l.shao}@philips.com

Abstract. In this paper we apply the Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) descriptor to the field of human action recognition. A video sequence is described as a collection of spatial-temporal words after the detection of space-time interest points and the description of the area around them. Our contribution has been in the description part, showing LBP-TOP to be a promising descriptor for human action classification purposes. We have also developed several extensions to the descriptor to enhance its performance in human action recognition, showing the method to be computationally efficient.

Keywords: Human action recognition, LBP-TOP, bag of words.

1 Introduction

Automatic categorization and localization of actions in video sequences has different applications, such as detecting activities in surveillance videos, indexing video sequences, organizing digital video library according to specified actions, etc. The challenge is how to obtain robust action recognition under variable illumination, background changes, camera motion and zooming, viewpoint changes and partial occlusions, geometric and photometric variations of objects and intra-class differences.

There are two main approaches: holistic and part-based representations. Holistic representations focus on the whole human body trying to search characteristics such as contours or pose. Usually holistic methods, which focus on the contours of a person, do not consider the human body as being composed of body parts but consider the whole form of human body in the analyzed frame. Efros et al. [1] use cross-correlation between optical flow descriptors and Shechtman et al. [2] use similarity between space-time volumes which allows finding similar dynamic behaviors and actions. Motion and trajectories are also commonly used features for recognizing human actions, e.g. Ali et al. [3] use trajectories of hands, feet and body. Holistic methods may depend on the recording conditions such as position of the pattern in the frame, spatial resolution, relative motion with respect to the camera and can be influenced by variations in the background and by occlusions. These problems can be solved in principle by external mechanisms (e.g. spatial segmentation, camera stabilization, tracking etc.), but such mechanisms might be unstable in complex situations and require more computational demand.

Part-based representations typically search for Space-Time Interest Points (STIPs) in the video, apply a robust description of the area around them and create a model

X. Jiang and N. Petkov (Eds.): CAIP 2009, LNCS 5702, pp. 740–747, 2009.

based on independent features (Bag of Words) or a model that can also contain structural information. These methods do not require tracking and stabilization and are often more resistant to cluttering, as only few parts may be occluded. Different methods for detecting STIPs have been proposed, such as [11], [12]. The resulting features often reflect interesting patterns that can be used for a compact representation of video data as well as for interpretation of spatio-temporal events.

The paper is organized as follows. In section 2 we provide the methodology adopted for classification and in Section 3 we provide an introduction to the LBP and LBP-TOP descriptors on 3D data. Experimental results on human action recognition are shown and evaluated in Section 4. Finally, we conclude in Section 5.

2 Methodology

In the following sections we describe our algorithm in detail. In Section 2.1 we explain the classification scheme of our algorithm. In Section 2.2 we give a brief description about the detection of STIPs and the feature description method is introduced in Section 2.3. Section 2.4 explains the classifier used.

2.1 Bag of Words Classification

The methodology we adopt is a Bag of Words classification model [11]. As a first step, space-time interest points are detected using a separable linear filter and small video patches (named cuboids) are extracted from each interest point. They represent the local information used to learn and recognize the different human actions. Each cuboid is described using the LBP-TOP descriptor. The result is a sparse representation of the video sequence as cuboid descriptors. Having obtained all these data for the training set, a visual vocabulary is built by clustering using the k-means algorithm. The center of each cluster is defined as a spatial-temporal ‘word’ of which length depends on the length of the descriptor adopted. Each feature description is successively assigned to the closest (we use Euclidean distance) vocabulary word and a histogram of spatial-temporal word occurrence in the entire video is computed. Thus, each video is represented as a collection of spatial-temporal words from the codebook in the form of a histogram. For classification, we use non linear Support Vector Machines (SVM). As the algorithm has random components, such as the clustering phase, any experiment result reported is averaged over 20 runs. The entire methodology used is shown in Fig. 1.

2.2 Feature Detection

Several spatio-temporal feature detection methods have been developed recently and among them we chose Dollar’s feature detector [11] because of its simplicity, fastness and because it generally produces a high number of responses. The detector is based on a set of separable linear filters which treats the spatial and temporal dimensions in different ways. A 2D Gaussian kernel is applied only along the spatial dimensions (parameter σ to be set), while a quadrature pair of 1D Gabor filters are applied only temporally (parameter τ to be set). This method responds to local regions which exhibit complex motion patterns, including space-time corners. For more implementation details, please refer to [11] as the feature detection part is beyond the scope of this paper.

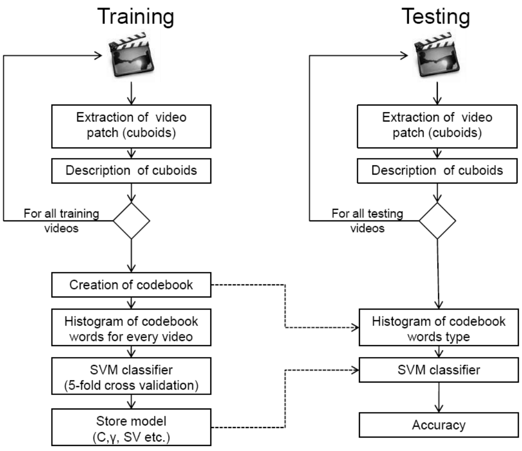
Training
Testing
Extraction of video patch (cuboids)
Description of cuboids
For all training videos
Creation of codebook
Histogram of codebook words for every video
SVM classifier (5-fold cross validation)
Store model (C,γ, SV etc.)
Extraction of video patch (cuboids)
Description of cuboids
For all testing videos
Histogram of codebook words type
SVM classifier
Accuracy

Fig. 1. Methodology adopted for action recognition

2.3 Feature Description

Once the cuboid is extracted, it is described using the LBP-TOP descriptor, which is an extension of LBP operator into the temporal domain. LBP has originally been proposed for texture analysis and classification [4]. Recently, it has been applied on face recognition [5] and facial expression recognition [6], [7]. While the original LBP was only designed for static images, LBP-TOP has been used for dynamic textures and facial expression recognition [8]. As a video sequence can not only be seen as the usual stack of XY planes in the temporal axis, but also as a stack of YT planes on X axis and as a stack of XT planes on Y axis, we prove that a cuboid can be successfully described with LBP-TOP for action recognition purposes.

2.4 Classification

Each video sequence is described as a histogram of space-time words occurrence which represents its signature. The dimension of the signature is equal to the size of the codebook and is given as input to the classifier (see Fig. 1). We chose to use non linear Support Vector Machines (SVM) with rbf kernel and the library libSVM [14] was adopted. The best parameters C and γ were chosen doing a 5-fold cross validation in a grid approach on the training data and one against one approach has been used for multi-class classification.

3 LBP-TOP and Its Extensions

The Local Binary Pattern (LBP) operator labels the pixels of an image by thresholding a circular neighborhood region [4]. The $LBP_{P,R}$ operator produces 2^P different output values, corresponding to the 2^P different binary patterns that can be formed by the *P* pixels in the neighbor set. The derived binary numbers encode local primitives such as curved edges, spots, flat areas etc. After the computation of the LBP for the whole image, an occurrence histogram of the labels is used as feature. It contains information about the distribution of local micro-patterns over the whole image and represents a statistical description of image characteristics. This descriptor has been proved to be successful in face recognition [5]. For more details about LBP operator, please refer to [4], [5], [6], [7]. Recently, LBP has been modified in order to be used in the context of dynamic texture description and recognition and for facial expression analysis [8]. LBP-TOP computes the LBP from Three Orthogonal Planes, denoted as XY-LBP, XT-LBP and YT-LBP. The operator is expressed as .

$$LBP-TOP_{P_{XY},P_{XT},P_{YT},R_X,R_Y,R_T} \,. \tag{1}$$

where the notation (P_{XY}, P_{XT}, P_{YT}, R_X, R_Y, R_T) denotes a neighborhood of *P* points equally sampled on a circle of radius *R* on XY, XT and YT planes respectively. The statistics on the three different planes are computed and then concatenated into a single histogram. The resulting feature vector is of $3{\cdot}2^P$ length. Fig. 2 illustrates the construction of the LBP-TOP descriptor. In such a scheme, LBP encodes appearance and motion in three directions, incorporating spatial information in XY-LBP and spatial temporal co-occurrence statistics in XT-LBP and YT-LBP.

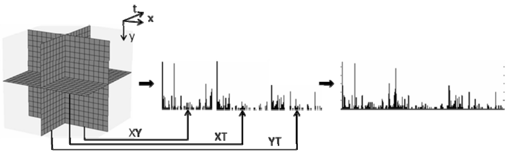

Fig. 2. LBP-TOP methodology

In our implementation, LBP-TOP is applied on each cuboid, as shown in Fig. 3, where XY, XT and YT planes are the central slices of it as can be seen in Fig. 4. Kellokumpu et al [8] have recently used LBP-TOP for human detection and activity description. However, their approach is based on background subtraction using LBP-TOP and a bounding volume has to be built around the area of motion. Their method can be categorize as holistic, since no space-time interest points have to be detected and differs from our part-based approach.

3.1 Modifications on LBP-TOP

As we described previously, the original LBP-TOP descriptor is the computation of LBP on the gray-level values of 3 orthogonal slices of each cuboid. We propose to extend the computation of LBP to 9 slices, 3 for each axis. Therefore, on the XY dimension we have the original XY plane (centered in the middle of the cuboid) plus other two XY planes located at 1/4 and 3/4 of the cuboid's length. The same is done for XT and YT dimensions. We named this method as Extended LBP-TOP. In this manner, more dynamic information in the cuboid can be extracted, as the 3 slices in one axis capture the motion at different times. We also exploit more information from the cuboid, dealing with 6 slices on each axis, located from 2/8 until 7/8 of the cuboid's length for each axis. In this case, a dimensionality reduction technique has to be applied since the final dimension of the descriptor vector would be too high.

Another modification we introduced is the computation of LBP operator on gradient images. The gradient image contains information about the rapidity of pixel intensity changes along a specific direction, has large magnitude values at edges and it can further increment LBP operator's performances, since LBP encodes local primitives such as curved edges, spots, flat areas etc. For each cuboid, the brightness gradient is calculated along *x*, *y* and *t* directions, and the resulting 3 cuboids containing specific gradient information are summed in absolute values. Before computing the image gradients, the cuboid is slightly smoothed with a Gaussian filter in order to reduce noise. LBP-TOP is then performed on the gradient cuboid and we name this method Gradient LBP-TOP. The Extended LBP-TOP can be applied on the gradient cuboid and we named this method as Extended Gradient LBP-TOP.

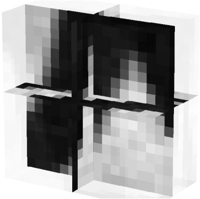

Fig. 3. Cuboid with XY, XT and YT planes

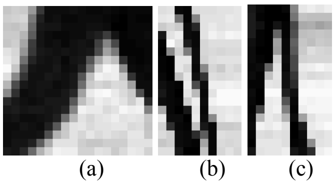

Fig. 4. Extracted XY *(a)*, XT *(b)* and YT *(c)* planes from the cuboid of Fig. 3

4 Experimental Results

For our action recognition experiments, we chose to use the KTH human action dataset [10]. This dataset contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. Each action class is performed several times by 25 subjects in different scenarios of outdoor and indoor environment. The camera is not static and the videos contain scale changes. In total, the dataset contains 600 sequences. We divide the dataset into two parts: 16 people for training and 9 people for testing, as it has been done in [10] and in [13]. We limit the length of all video sequences to 300 frames.

We extract the space-time interest points and describe the corresponding cuboids with the procedure described in Sections 2.2 and 2.3. The detector parameters are set to σ=2.8 and τ=1.6, which gave better results in our evaluations, and 80 STIPs were detected for each sequence. The original LBP-TOP and the Extended LBP-TOP are computed on the original cuboid or on the gradient cuboid. The number of clusters used to build the codebook is chosen to maximize the classification accuracy on the testing data and best values have been achieved using 1000 visual-words.

The accuracy results for LBP-TOP with different parameters *R* and *P* are shown in Table 1. The notation of parameters is as illustrated in Equation (1). Better classification accuracy has been obtained with the parameter *P* greater than 6 and radius *R* equal to 2. The performance is generally slightly decreasing as the radius *R* is getting bigger, while it is increasing as the number of neighbors *P* is increased. This could be explained as more neighbors permit to take more information into account. However, the drawback is a higher computational cost and a higher dimensionality of the feature vector.

Table 1. Accuracy for different parameter values of LBP-TOP$_{P,P,P,R,R,R}$

		Neighbors (P)			
		4	6	8	10
Radius (R)	2	71.81%	85.65%	86.25%	86.32%
	3	84.54%	85.18%	85.12%	86.69%
	4	81.34%	85.12%	85.46%	83.82%

LBP-TOP$_{8,8,8,2,2,2}$ produces a 768 vector length, while LBP-TOP$_{10,10,10,2,2,2}$ has a descriptor dimension of 3072. The final descriptor of LBP-TOP$_{12,12,12,2,2,2}$ will be 12288 vector lengths. The use of uniform LBP operator decreases the performance results compared with the original operator, since less information is kept into account (see Table 2). Multiresolution LBP operator has also been tested, but the gain in performances is not considerable with the increase of the descriptor length and computational cost. The time calculated in the following tables is measured on a computer equipped with a 3 Ghz Pentium 4 CPU and 3 Gb RAM.

We choose to use LBP-TOP$_{8,8,8,2,2,2}$ for the following experiments as it is computationally more efficient and the accuracy is among the highest. As dimensionality reduction technique, we used Principal Component Analysis (PCA) and set the final dimension to 100.

In Table 2, the Extended LBP-TOP is evaluated and different number of slices is taken into account. As we can see, the Extended LBP-TOP descriptor performs better than the original one, since more information is taken into consideration at different times in XY planes and at different locations in the XT and YT planes. Although best result is obtained with 6 slices on each axis, the computational time is almost double than the Extended LBP-TOP version with 3 slices; because of this issue, in the following we are computing the Extended version on only 3 slices for each axis.

Table 3 is a summary of best results achieved for different enhancement of LBP-TOP. The usage of LBP-TOP applied to the gradient cuboid gives better results compared with the original one. The information extracted from the gradient calculated along *x*, *y* and *t* directions and combined into the gradient cuboid permits to have a better performance for LBP-TOP in the description of actions. Moreover, a slight increase in performances can be achieved by applying the Extended LBP-TOP on the gradient cuboids. The number of support vectors calculated by SVM is, in all methods tested, about 360. As a feature reduction method, we applied PCA and show that the classification accuracy is only decreased slightly in Extended LBP-TOP, while slightly increasing in the Extended Gradient LBP-TOP.

Table 2. Uniform and Extended LBP-TOP

Method	Accuracy	Descriptor length	Computational time (s)
LBP-TOP$_{8,8,8,2,2,2}$	86.25 %	768	0.0139
Uniform LBP-TOP$_{8,8,8,2,2,2}$	81.78 %	177	0.0243
Extended LBP-TOP$_{8,8,8,2,2,2}$ (3 slices on each axis)	88.19 %	2304	0.0314
Extended LBP-TOP$_{8,8,8,2,2,2}$ (3 slices on each axis) + PCA	87.87 %	100	0.0319
Extended LBP-TOP$_{8,8,8,2,2,2}$ (6 slices on each axis) + PCA	88.38 %	100	0.0630

Table 3. Accuracy for different LBP-TOP methods

Method	Accuracy	Descriptor length	Computational time (s)
Ext LBP-TOP$_{8,8,8,2,2,2}$	88.19 %	2304	0.0314
Ext LBP-TOP$_{8,8,8,2,2,2}$ + PCA	87.87 %	100	0.0319
Grad LBP-TOP$_{8,8,8,2,2,2}$	90.07 %	768	0.0788
Ext Grad LBP-TOP$_{8,8,8,2,2,2}$	90.72 %	2304	0.0992
Ext Grad LBP-TOP$_{8,8,8,2,2,2}$ + PCA	91.25 %	100	0.1004
HOG-HOF	89.88 %	162	0.2820
HOG-HOF + PCA	89.28 %	100	0.2894

As a comparison, we evaluate Laptev's method [13] with the same framework as illustrated in Section 2. We use Laptev's code publicly available on his website and recently being updated with the latest settings used in [13]. The combination of Laptev's extraction method and Laptev's HOG-HOF descriptor make us reach an accuracy of 89.88%. The time for Laptev's HOG-HOF descriptor in Table 3 is referred to both extraction and description parts, as the description part cannot be computed regardless of the extraction part in Laptev's provided executable. Therefore, we expect the description part to be about half the time shown in the table. The computational time for HOG-HOF is affected by the choice of the threshold and we have chosen a suitable threshold to have 80 detected STIPs for this comparison. There is also to mention that Laptev's executable code is compiled in C environment, while our LBP-TOP implementation is compiled in Matlab environment. Similar performance to Laptev's is achieved using the Extended LBP-TOP descriptor which is almost 3 times computationally faster than the Extended Gradient LBP-TOP descriptor.

5 Conclusion

In this paper, we have applied LBP-TOP as a descriptor of small video-patches used in a part-based approach for human action recognition and shown LBP-TOP to be suitable for the description of cuboids containing information about human actions. We have extended LBP-TOP considering the action at three different frames in XY plane and at different views in XT and YT planes. Furthermore, we applied LBP-TOP to gradient images. We have also shown that the performance of descriptor is quite stable when the PCA is applied. Regarding computational time, the Extended Gradient LBP-TOP descriptor, compared with HOG-HOF, is computationally more efficient and permits to reach better accuracy in our framework. The experimental results reveal that LBP-TOP and its modifications tend to be good candidates for human action description and recognition. The best accuracy has been obtained by using the Extended Gradient LBP-TOP$_{8,8,8,2,2,2}$ with PCA.

References

1. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Proceedings of Ninth IEEE International Conference on Computer Vision, vol. 2, pp. 726–733 (2003)
2. Shechtman, E., Irani, M.: Space-time behavior based correlation. In: IEEE Computer Society Conference on CVPR, June 20-25, vol. 1, pp. 405–412 (2005)
3. Ali, S., Basharat, A., Shah, M.: Chaotic invariants for human action recognition. In: Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 1–8 (2007)
4. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 971–987 (2002)
5. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J.G. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
6. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Description with Local Binary Patterns: Application to Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12), 2037–2041 (2006)
7. Shan, C., et al.: Facial Expression recognition based on Local Binary Patterns: A comprehensive study. Image and Vision Computing (2008)
8. Zhao, G., Pietikäinen, M.: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6), 915–928 (2007)
9. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Human Activity Recognition Using a Dynamic Texture Based Method. In: British Machine Vision Conference (2008)
10. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition, August 2004, vol. 3, pp. 32–36 (2004)
11. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.J.: Behavior recognition via sparse spatio-temporal features. In: Proc. of ICCV Int. work-shop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VSPETS), pp. 65–72 (2005)
12. Laptev, I.: On space-time interest points. International Journal of Computer Vision (IJCV) 64(2-3), 107–123 (2005)
13. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
14. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`