Feature Detector and Descriptor Evaluation in Human Action Recognition

Ling Shao
[1]Department of Electronic & Electrical Engineering
The University of Sheffield
Sheffield, UK
[2]Shenzhen Institute of Advanced Integration Technology
CAS/CUHK, China
ling.shao@sheffield.ac.uk

Riccardo Mattivi
Department of Information Engineering and Computer Science,
University of Trento
Via Sommarive 14 I-38100
Povo (TN) - Italy
rmattivi@disi.unitn.it

ABSTRACT

In this paper, we evaluate and compare different feature detection and feature description methods for part-based approaches in human action recognition. Different methods have been proposed in the literature for both feature detection of space-time interest points and description of local video patches. It is however unclear which method performs better in the field of human action recognition. We compare, in the feature detection section, Dollar's method [18], Laptev's method [22], a bank of 3D-Gabor filters [6] and a method based on Space-Time Differences of Gaussians. We also compare and evaluate different descriptors such as Gradient [18], HOG-HOF [22], 3D SIFT [24] and an enhanced version of LBP-TOP [15]. We show the combination of Dollar's detection method and the improved LBP-TOP descriptor to be computationally efficient and to reach the best recognition accuracy on the KTH database.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.5.4 [**Pattern Recognition**]: Applications;

General Terms

Algorithms, Performance, Design, Experimentation, Theory.

Keywords

Human Action Recognition, LBP-TOP, Bag of Words, Feature Detectors, Feature Descriptors.

1. INTRODUCTION

Automatic categorization and localization of actions in video sequences has different applications, such as detecting activities and behaviors 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, partial occlusions, geometric and photometric variations of objects and intra-class differences.

In the literature, two main approaches for human action recognition are used: holistic and part-based representations. The holistic representations focus on the whole body of the person, trying to search for characteristics such as contours or pose. These methods, which focus on the contours of a person, usually consider the whole form of human body in the analyzed frame. Efros et al. [1] use cross-correlation between optical flow descriptors while Shechtman et al. [2] measure the similarity between space-time volumes. This allows us to find similar dynamic behaviors and actions. Ali et al. [3] use trajectories of hands, feet and body. The performance of holistic methods may depend on recording conditions such as spatial resolution, position of the pattern in the frame and relative motion with respect to the camera. Moreover, global image measurements can be influenced by motions of multiple objects, variations in the background and occlusions.

Part-based representation consist of two steps: the feature detection phase, in which space-time interest points in the video are searched; the feature description phase, in which a robust description of the area around them is computed and a model based on independent features (Bag of Words) or a model that can also contain structural information is built. These methods do not require tracking and stabilization and can be more resistant to cluttering as only a few parts may be occluded. The Bag of Words classification method has been recently applied to action recognition [18]. Different methods for detecting space-time interest points have been proposed, such as Dollar's method [18], Laptev's method [21], Ning's method [6], together with several video patch descriptors, such as Gradient [18], HOG-HOF [22], 3D SIFT [24]. 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. In the 2D domain, Mikolajczyk et al. have done a comprehensive evaluation and comparison of affine region detectors [4] and a performance evaluation of local descriptors [5].

However, no evaluation and comparison have been carried out in the space-time domain for different detection and description methods. The aim of this paper is to compare different detection methods and different description methods for spatio-temporal features in the field of human action recognition. Our comparison is different to that done in [23], because we use different feature

CIVR 2010, July 5–7, 2010, Xi'an, China.
Copyright ©2010 ACM 978-1-4503-0117-6/10/07... $10.00.

detectors and descriptors and focus on the strength of spatio-temporal interest points for action recognition.

In Section 2, we describe the methodology used for the experimental setup. Section 3 and Section 4 summarize the feature detection and feature description methods used in our comparison, respectively. In Section 5, we present the experimental results. Finally, we discuss the results and conclude in Section 6.

2. METHODOLOGY

The methodology we adopt is a Bag of Words classification model [18]. Space-time interest points are detected, as a first step, using one feature detection method 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 one feature description method. 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 (using Euclidean distance) vocabulary word. A histogram of spatial-temporal word occurrence in the entire video is then 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) and k Nearest Neighbors (kNN) classifier. As the algorithm has a random component, the clustering phase, any experiment result reported is averaged over 20 runs. In Figure 1 the entire methodology is shown.

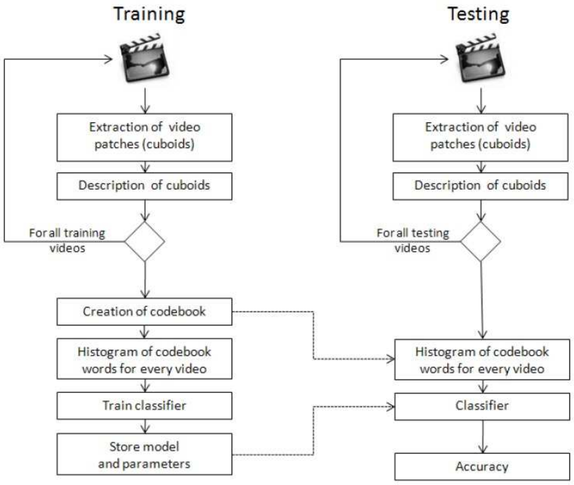
Training
Testing
Extraction of video patches (cuboids)
Description of cuboids
For all training videos
Creation of codebook
Histogram of codebook words for every video
Train classifier
Store model and parameters
Extraction of video patches (cuboids)
Description of cuboids
For all testing videos
Histogram of codebook words for every video
Classifier
Accuracy

Figure 1: Methodology for human action recognition.

3. FEATURE DETECTION METHODS

In this section, various space-time interest point detectors are being described. The experimental results on action recognition are shown in Section 5 and the settings of parameters are discussed in sub-section 5.1.1.

3.1 Dollar's Method

The detector proposed by Dollar et al. [18] is based on a set of separable linear filters which treats the spatial and temporal dimensions in different ways. The response function is given by

$$R = (I * g * h_{even})^2 + (I * g * h_{odd})^2 \qquad (1)$$

where g(x,y,σ) is a 2D Gaussian kernel, applied only along the spatial dimensions, and h_{even} and h_{odd} are a quadrature pair of 1D Gabor filters applied only temporally. They are defined as $h_{even}(t;\tau,\sigma) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{odd}(t;\tau,\sigma) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$. The authors suggested to keep ω=4/τ, and the number of parameters to be set in the response function *R* is reduced to two. The parameters σ and τ correspond roughly to the spatial and temporal scales of the detector, respectively. This method responds to local regions which exhibit complex motion patterns, including space-time corners. Regions undergoing motion with constant speed or without spatially distinguished features induce a low response. The space time interest points are detected around the local maxima of *R*.

3.2 Laptev's Method

Laptev and Lindeberg [19] presented a multiscale space-time interest point detector based on the idea of Harris and Förstner interest point operators [20]. They detect local structures in space-time where the image values have significant local variations in both dimensions. Gradients can be found not only along x and y, but also along t and spatio-temporal corners are defined as regions where the local gradient vectors point in orthogonal directions spanning x, y and t. A spatio-temporal corner is therefore an image region containing a spatial corner whose velocity vector is in reversing direction. In our evaluation tests we use Laptev's code publicly available on his website and recently being updated with the latest settings used in [22].

3.3 Bank of 3D-Gabor Filters

This method uses a 3-Dimensional Gabor filter in order to localize interesting areas in the spatio-temporal dimension and it has been introduced by Ning et al. [6]. The Gabor filter is composed of two main components: the sinusoidal carrier and the Gaussian envelope. The original video sequence is convolved with a bank of 3D Gabor filters having different wavelengths of the underlying cosine and different orientations in space. The Gabor filter has been introduced in the field of human action recognition as it exhibits many common properties to mammalian cortical cells, such as spatial localization, orientation selectivity and spatial frequency characterization. A bank of Gabor filters has also been used in object recognition [7].

3.4 Space-Time DoG

In the context of local image features, Lowe [7] developed a method for extracting distinctive scale-invariant features by finding extrema in Differences of Gaussians (DoG). Inspired by Lowe's detection of stable keypoint locations in scale-space, Cheung et al. [10] propose a natural extension of it into the third dimension for biomedical image processing purposes. We applied this method in order to find stable keypoint locations not only in scale-space, but also in time.

The initial video sequence is incrementally convolved with Gaussians to produce images separated by a constant factor k in scale-space-time. To obtain candidate feature points, each sample point is compared to its 26 neighbors at time t, t-1 and t+1 in the current scale and its 27 neighbors at time t, t-1 and t+1 in the scale above and below, respectively. The space-time interest point is selected only if its response is larger than all of these neighbors. After the detection of scale-space extrema, the edge-like features are discarded.

4. FEATURE DESCRIPTION METHODS

After the localization of a space-time interest point in the video sequence, a video patch is extracted around the interest point location and described using one of the description methods. In this section, we give a brief overview for each description method and in sub-section 5.1.2 the details of the parameter settings.

4.1 Gradient

The Gradient descriptor was introduced by Dollar et al. [18] and is obtained calculating the brightness gradients of the cuboid along x, y and t directions. The spatial-temporal cube is first smoothed at different scales before computing the image gradients. The computed gradient values are concatenated to form a vector. This vector, which represents the descriptor, is then projected to a lower dimensional space using the principal component analysis (PCA) dimensionality reduction technique.

Dollar et al. proved the gradient descriptor to perform better then normalized pixel values or optical flow. Moreover, the simple concatenation of values in a vector was proved to perform better than the N-Dimensional Histograms and local N-Dimensional Histograms [18].

4.2 HOG-HOF

Laptev et al. [22] use the Histogram of Oriented Gradients (HOG) as cuboid descriptor as local object appearance and shape can be characterized by the distribution of local intensity gradients. HOG descriptor is implemented by dividing the cuboid into small space-time regions and accumulating a local 1-D histogram of gradient directions over the pixels of each sub-region. The combined histogram entries form the representation. Another approach used by Laptev is Histogram of Optic Flow (HOF). The idea is the same as the previous descriptor HOG, with the only difference that the histogram of optic flow is computed for each sub-region. Laptev proved the combination of HOG and HOF, named HOG-HOF, to perform better than each separate method.

4.3 3D SIFT

The 3-Dimensional Scale-Invariant Feature Transform (3D SIFT) descriptor is an extension of SIFT into 3-dimensional space developed by Scovanner el at. [24]. The underlying idea is similar to HOG because both methods are encoding gradient characteristics in 3D space. The gradient magnitude and orientation in 3D SIFT descriptor are given by

$$m_{3D}(x, y, t,) = \sqrt{L_x^2 + L_y^2 + L_t^2} \tag{2}$$

$$\theta(x, y, t) = \tan^{-1}(L_y / L_x) \tag{3}$$

$$\phi(x, y, t) = \tan^{-1}\left(\frac{L_t}{\sqrt{L_x^2 + L_y^2}}\right) \tag{4}$$

In this manner, each pixel has two values which represent the direction of the gradient in three dimensions: θ encodes the angle in the 2D gradient direction, while Φ encodes the angle away from the 2D gradient direction. A sub-histogram is then created by sampling the sub-regions surrounding the interest point. For each 3D sub-region the orientations are accumulated into a histogram and the final descriptor is a vectorization of the sub-histograms. Fig. 3 illustrates the calculation of 3D SIFT.

4.4 LBP-TOP

Local Binary Pattern computed on Three Orthogonal Planes (LBP-TOP) has been developed by Zhao et al. [15] and has been successfully used for dynamic texture description and recognition and has been applied also on facial expression analysis. LBP-TOP algorithm extracts the LBP code from three orthogonal planes, encoding appearance and motion in three directions. The spatial information is encoded in XY plane and the spatial temporal co-occurrence statistics are encoded in XT and YT planes. Given a pixel at a certain location, the original LBP operator [12] can be expressed in decimal form as

$$LBP_{P,R} = \sum s(g_p - g_c)2^p \tag{5}$$

where the notation (*P,R*) denotes a neighborhood of *P* points equally sampled on a circle of radius *R*, g_c is the gray-level value of the central pixel and g_p are the *P* gray-level values of the sampled points in the neighborhood, *s(x)* is 1 if *x*≥0 and 0 otherwise. For more details about LBP, please refer to [10][12][13]. Another version of LBP has been recently developed by Heikkila et al. [16] where the pixels are compared in a different manner, giving a descriptor whose length is 16 times shorter

$$CS - LBP_{P,R} = \sum_{i=0}^{(P/2)-1} s(g_i - g_{i+(P/2)})2^i \tag{6}$$

The statistics on the three different planes of LBP-TOP descriptor are computed histogramming the LBP (or CS-LBP) values and concatenating them into a single histogram.

In [30], we propose to extend the computation of LBP and CS-LBP to 9 slices, 3 for each axis. We name this method as Extended LBP-TOP and Extended CSLBP-TOP, respectively. Using the extended version, more dynamic information inside the cuboid can be extracted. On XY plane, for instance, 3 slices capture the motion at different times. Another modification we introduced is the computation of LBP and CS-LBP operator on gradient images of the orthogonal slices. The gradient image contains information about the rapidity of pixel intensity changes along a specific direction and has large magnitude values at edges. Gradient image can further increment LBP operators' performance, since LBP encodes local primitives such as curved edges, spots, flat areas etc [12]. LBP-TOP is then performed on the gradient cuboid and we name this method as Gradient LBP-TOP and Gradient CSLBP-TOP. The Extended LBP-TOP and Extended CSLBP-TOP methods can be applied on the gradient cuboid.

5. RESULTS

5.1 Experimental Setup

In the following, we explain implementation details and parameters used in our comparison. Moreover, we discuss the evaluation criteria and the classifiers used.

5.1.1 Implementation details – detection methods

The parameters that can be set during the detection part are the number of space-time interest points and the intrinsic parameters for each method.

Dollar's separable filters: the parameters need to be set are σ for the 2D Gaussian smoothing kernel and τ for the quadrature pair of 1D Gabor filters applied temporally. The parameter ω of the underline frequency of cosine in the Gabor filter is related to τ by the following relation $\omega=4/\tau$. For simplicity, we run the detector using only one scale with parameters $\sigma=2.8$ and $\tau=1.6$, which gave better results in our evaluations.

Laptev's space-time corners: we used Laptev's latest implementation based on a multiscale approach which does not do scale selection [22]. The code is publicy available on Laptev's website. We set the detector with the suggested parameters: 3 pyramid levels and a patch size factor equal to 5.

3D-Gabor: rotations in xy plane are applied and three discrete values for θ = [-90 0 90] degrees and three different values for λ are used. Nine different 3D Gabor filters are then built and convolved with the original video sequence.

Space-Time DoG: best performances were obtained adopting k=5 scales for each of s=3 octaves. This setup gives 15 convolutions to be computed.

5.1.2 Implementation details – description methods

Gradient descriptor is simply a concatenation of gradient values along the three dimensions. The size of the vector is equal to the number of pixels in the cube times the number of smoothing scales times the number of gradient directions. In our implementation, we smoothed the cuboid only once. The vector's length corresponds therefore to three times the volume of the cuboid, i.e. a cuboid of 19x19x11 pixels gives a vector of 11913 dimensions. The vector is then reduced to 100 dimensions with PCA.

HOG-HOF is a concatenation of HOG descriptor (72 vector length) and HOF descriptor (90 vector length). The implementation is set by Laptev's code.

3D SIFT is a descriptor whose length depends on the number of sub-histograms and the number of bins used to represent θ and φ angles. The implementation used is taken from Scovanner's publicly available code with the suggested parameters (which are slightly changed from what is described in [24]) giving a vector length of 640 dimensions.

LBP-TOP: original implementation [15] gives a descriptor whose length depends exponentially on the number of neighbors; since the $LBP_{P,R}$ operator produces 2^P different output values, LBP-TOP's final descriptor is of $3\cdot2^P$ vector length, while the Extended version produces a $9\cdot2^P$ vector length. CS-LBP operator produces $2^{P/2}$ different output values, therefore CSLBP-TOP's and Extended CSLBP-TOP's final descriptors are of $3\cdot2^{P/2}$ and $9\cdot2^{(P/2)}$ vector lengths, respectively. If PCA is applied, the dimension is reduced to 100. If not specified, LBP is computed with parameters P=8 and R=2.

5.1.3 Dataset

For our action recognition experiments, we chose to use the KTH human action dataset [17]. It contains six types of human actions: *walking, jogging, running, boxing, handwaving and handclapping*. 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 video sequences 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 [22] [27]. We limit the length of all video sequences to the first 300 frames.

5.1.4 Classification

Each video sequence is described as a histogram of space-time words occurrence and represents the signature of each video. The dimension of the signature is equal to the size of the codebook and is given as input to the classifier. We chose to use non-linear Support Vector Machines (SVM) with rbf kernel and the library libSVM [28] was used. 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. We also use a 1-NN classifier with the χ^2 distance, as suggested in [18].

5.2 Feature Detection Methods

In this section, we show the classification accuracy for different detection methods as the number of cuboids extracted from every video is increasing. We keep fixed the description method and we chose to use the Gradient descriptor, as it has been proved by Dollar et al. [18] to give better classification accuracy compared with normalized pixel values, brightness gradient and windowed optical flow. Moreover, it is computationally efficient. The size of each cuboid is kept constant for all methods (19x19x11). We chose to have a codebook of k=1000 visual words, as it is shown in section 5.3 to give overall better results.

As we can see in Figure 2 and Figure 3, Dollar's method performs better among all tested methods. We can further notice that for all methods the performance is increasing as the number of cuboids increases. This is due to the fact that a high number of features is required for Bag of Words classification, as it is proved, for example, in 2D natural scene classification [9]. Dollar's detection method gives much better recognition accuracy than Laptev's method when a small number of cuboids are extracted from the original video sequences. This could be explained as Dollar's method is more accurate in finding interest point locations suitable for better distinguishing among the different action types and as space-time corners used in Laptev's method are quite rare. By setting a lower threshold, which means taking a higher number of cuboids, Laptev's method is increasing the performance, but the accuracy results are always worse than Dollar's. Laptev's code uses a multiscale approach, that could not be so useful in KTH database as the videos are, in most cases, taken from the same distance.

The bank of 3D Gabor filters is shown to have the classification accuracy lower than the previous two methods. This fact could be explained as the 3D Gabor filter extracts interest points that are not so relevant for describing the motion, since complex regions undergoing limited motion could also give high response for the 3D Gabor filter. However, as the number of cuboids increases, the performance of 3D Gabor filter reaches that of Laptev's method. Space-Time DoG is also shown to have classification accuracy lower than the previous methods, but the performance is increasing again as the number of cuboid increases.

Regarding computational time, in Table 1 a comparison of different detection methods is shown. The time is referred as an average over 10 runs for the detection of 80 space-time interest points of a video sequence whose length is 300 frames. The time calculated in this table and in the following Table 2 is measured on a computer equipped with a 3 Ghz Pentium 4 CPU and 3 Gb RAM.

In Laptev's provided executable, the program is computing both detection and description parts at the same time. Therefore, computational time of Laptev's method shown here is longer than it should be for only the detection part, and we expect this part to be about half the time shown in the table.

As we can see, Dollar's method is the most efficient followed by Laptev's, disregarding the previously computational time mentioned issue. The results show that 3D-Gabor filter and Space-Time DoG are less accurate than Dollar's or Laptev's methods. Moreover, they are computationally much more expensive.

Table 1: Computational Time for Different Feature Detection Methods

Detection method	Environment	Comp. time (s)
Dollar	Matlab	7.61
Laptev	Optimized C	41.96
3D-Gabor	Matlab	84.53
Space-Time DoG	Matlab	~380

5.3 Feature Description Methods

In this section we present and discuss the experimental results for different descriptors. We chose to use Dollar's feature detection method for all our tests because of its simplicity, fastness and overall better accuracy. However, in the case of HOG-HOF descriptor, Laptev's feature detection method is used because, in Laptev's provided executable, the description part cannot be computed regardless of the detection part. Moreover, we wanted to have a reference to check for the correctness of our framework. The number of cuboids is fixed to 80 for all video sequences. This choice is due to previous results in section 5.2 and due to computational reasons. We chose to evaluate the descriptors' performance in their original dimension as well as in lower dimension after applying PCA, in order to have a fair comparison with descriptors of the same length. The eigenvalues of PCA are obtained using a subset of the training set.

In Figure 4 and Figure 5, results for different LBP-TOP description methods are shown. The best classification accuracy has been obtained with the Extended Gradient LBP-TOP. If PCA is applied, the performance does not decrease much. In general, CSLBP-TOP performs worse than the original LBP-TOP in the field of human action recognition. Only if the number of neighbors are increased (i.e. P=12), Extended CSLBP-TOP is almost reaching the performance of Extended LBP-TOP (P=8), as more spatial information is taken into account during the computation of CS-LBP. The Extended Gradient CSLBP-TOP version is performing best among the descriptors based on CS-LBP operator, reaching and overcoming with p=12 the performance of Gradient LBP-TOP.

In Fig. 6 and Fig. 7, other methods are evaluated and compared with the best promising descriptor based on LBP-TOP operator.

As the results show, best performance is obtained with a codebook size ranging from 750 to 1250 visual words for the majority of descriptors.

The performance of SVM and 1-NN is quite close; however, notice that the same descriptor behaves differently with different classifiers. This could be explained as the hyperplanes of SVM generalize too much in the division of different classes, while 1-NN keeps the division more defined, since all training samples are stored in the model. During the testing phase, a comparison with the first nearest neighbor is done and this permits to have a less generalized boundary between classes. For example, 3D SIFT has the best performance with SVM classifier, while the Extended Gradient LBP-TOP gives the best results using 1-NN. On average, 1-NN classifier with χ^2 distance gives better results in the case of Extended Gradient LBP-TOP with PCA applied, and the classification accuracy is 92.18% while with SVM 91.25%.

The correctness of our framework is proved by the fact that similar performance as in [22] is obtained using Laptev's code. The combination of Laptev's detection method and Laptev's HOG-HOF descriptor make us reach an accuracy of 89.88% with SVM. In [22], 91.8% of accuracy is obtained on KTH database, but different channel combinations for HOG and HOF are being used.

Regarding dimensionality reduction and descriptor of the same dimensionality, the performance is still similar to descriptors in their original dimension. In some cases, the descriptors reduced with PCA give better results for a small codebook size (e.g. 3D SIFT).

Table 2: Descriptor Length and Computational Time in Seconds for One Cuboid's Descriptor

Description method	Descriptor length	Comp time
Gradient + PCA	100	0.0060
3D SIFT	640	1.1180
LBP-TOP	768	0.0139
CS-LBP-TOP (p=10)	96	0.0115
Ext LBP-TOP	2304	0.0314
Ext Grad LBP-TOP	2304	0.0992
HOG-HOF	162	0.2820

In Table 2, computational time is shown for different description methods. The fastest method is Dollar's gradient, as the only operation to be done is a concatenation of pixels' gradient values. LBP-TOP is also quite fast, as the algorithm is mainly based on a comparison of pixel values. The Extended version is almost 3 times slower as LBP operator has to be computed on 9 slices. The Gradient version needs the computation of gradients along the three dimensions before LBP is applied, resulting in a larger computational time. CSLBP-TOP is faster than the original LBP-TOP as the number of comparison for each pixel is reduced by a factor of 2. 3D SIFT requires more time among all descriptors, since histograms have to be built for different values of θ and Φ.

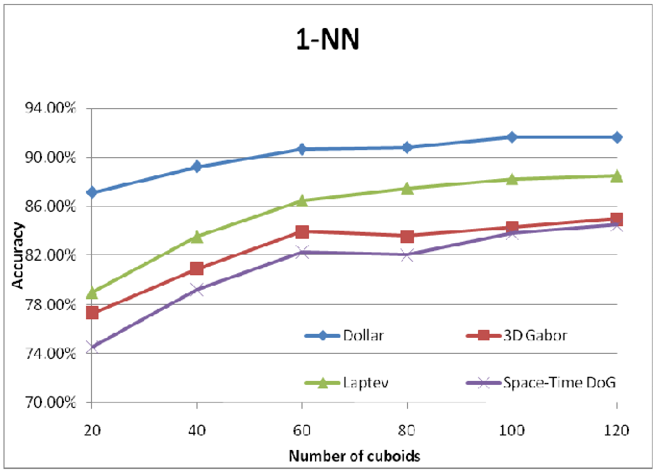

Figure 2. Cuboid's number vs. accuracy, 1-NN.

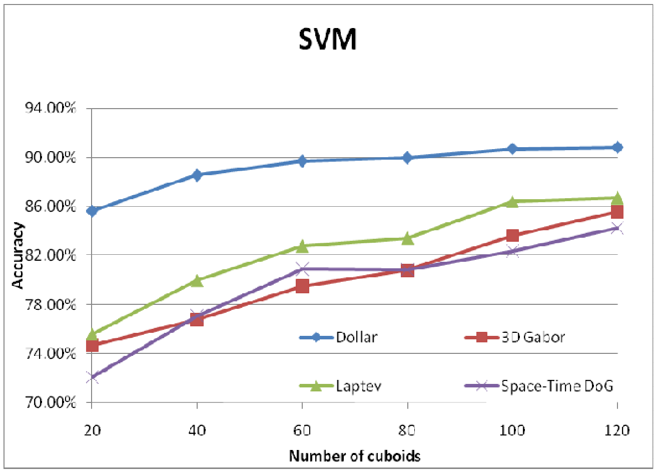

Figure 3. Cuboid's number vs. accuracy, SVM.

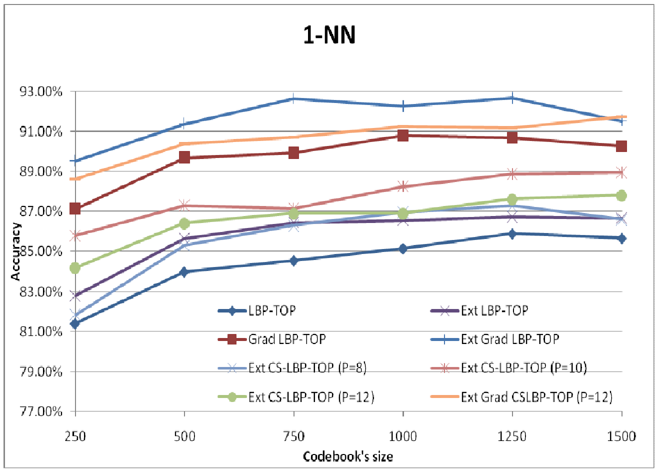

Figure 4. Codebook's size vs. accuracy for different LBP and CS-LBP based detection methods, 1-NN.

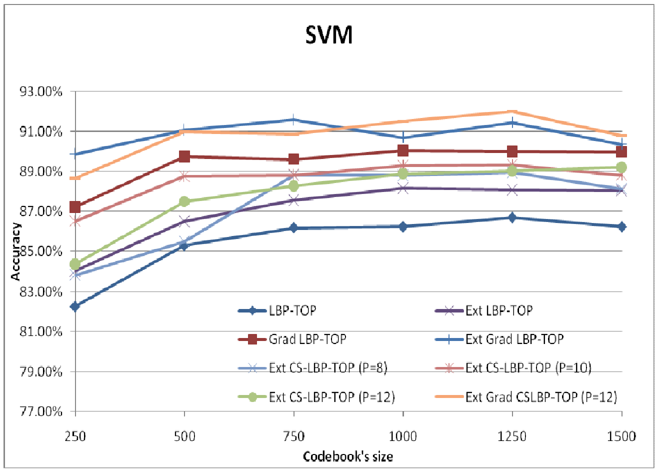

Figure 5. Codebook's size vs. accuracy for different LBP and CS-LBP based detection methods, SVM.

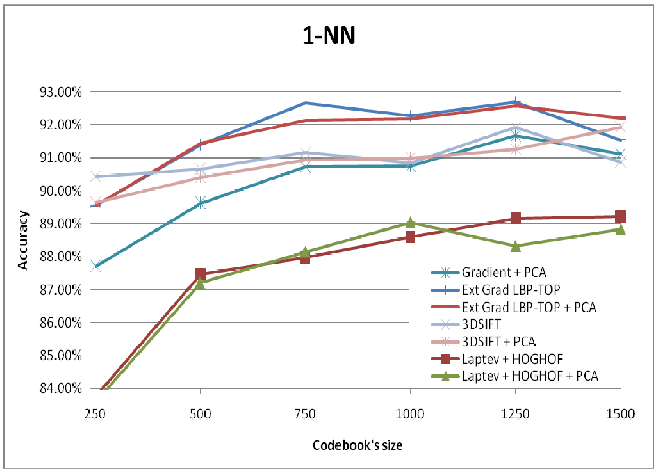

Figure 6. Codebook's size vs. accuracy for different detection methods, 1-NN.

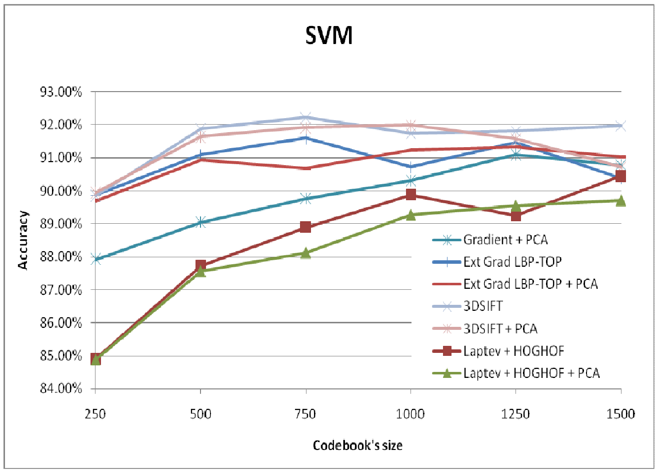

Figure 7. Codebook's size vs. accuracy for different detection methods, SVM.

All methods have been implemented in the Matlab environment, whereas Laptev's algorithm uses an optimized C code and, as previously mentioned, the time is for both detection and description parts. The computational time for HOG-HOF is affected by the choice of the threshold and we have chosen a suitable threshold to have a fixed amount of detected STIPs for this comparison. The usage of Extended Gradient LBP-TOP permits us to reach the best results on the KTH human action database by achieving 92.69% classification accuracy and 92.57% if PCA is applied using the 1-NN classifier with a codebook of 1250 'visual words'. In addition, we have also shown the enhanced LBP-TOP descriptor to be more efficient compared with 3D SIFT and HOG-HOF.

Generally, the Extended Gradient LBP-TOP reaches the top performance using 1-NN classifier, followed by 3D SIFT, Gradient and HOG-HOF. If SVM classifier is chosen, 3D SIFT gives the best performance, followed by Extended Gradient LBP-TOP, Gradient and HOG-HOF descriptors.

6. CONCLUSIONS

In this paper, we have presented an experimental evaluation of feature detection methods and feature description methods in the field of human action recognition. The objective is to find the best detection and description methods for future use in more challenging scenarios.

In the tests, Dollar's feature detection method proved to be faster, simpler, more precise and gives overall better performance, even though only one scale has been used.

The best descriptors are the Extended Gradient LBP-TOP and 3D SIFT, which give alternate performance using 1-NN or SVM classifiers. In reduced dimensions, they still give better performance compared with Gradient and HOG-HOF. One good property of the Extended Gradient LBP-TOP descriptor, compared with 3D SFIT, is its computational efficiency (approximately 9 times faster).

The best performance on the KTH database (92.69%) has been achieved using the combination of Dollar's detection method and the Extended Gradient LBP-TOP using 1-NN classifier with the χ^2 distance. We have also shown that the performance of descriptors is quite stable when PCA is applied and the final dimension of descriptors is set to 100.

In future work, similar experiments can be conducted for human action recognition on a more realistic and challenging database, such as the HOHA database 2 [29].

7. ACKNOWLEDGMENTS

The work described in this article was partially supported by Shenzhen Science Plan (Project No. SY200806300223A) and NSFC (Project No. 60903115).

8. REFERENCES

[1] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pp.726-733, vol.2, 13-16 Oct, 2003.

[2] E. Shechtman and M. Irani Space-time behavior based correlation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 405-412, vol. 1, 20-25 June, 2005.

[3] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pages 1-8, 2007.

[4] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, vol. 65(1/2): 43-72, 2005.

[5] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.27, no.10, pp.1615-1630, Oct. 2005

[6] H. Ning, Y. Hu, and T.S. Huang. Searching Human Behaviors using Spatial-Temporal words. In *Proceedings of IEEE International Conference on Image Processing*, vol.6, pp.VI -337-VI -340, September-October, 2007.

[7] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 994-1000, 2005.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91-110, 2004.

[9] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 524-531, 2005.

[10] W. Cheung and G. Hamarneh. N-sift: N-dimensional scale invariant feature transform for matching medical images. In *Proceedings of the* 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 720-723, 2007.

[11] T. Ojala, M. Pietikanen, and D. Harwood. A comparative study of texture measures with classification based on featured distribution. *Pattern Recognition*, 29(1):51–59, 1996.

[12] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24 (7), pp. 971–987, 2002.

[13] T. Ahonen, A. Hadid, M. Pietikäinen. Face recognition with local binary pattern. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2004.

[14] T. Ahonen, A. Hadid, and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.12, pp.2037-2041, December 2006.

[15] G. Zhao and M. Pietikäinen. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.29, no.6, pp.915-928, June 2007.

[16] M. Heikkila, M. Pietikainen, and C.Schmid. Description of interest regions with center-symmetric local binary patterns. In *Proceedings of the 5th Indian Conference on Computer Vision, Graphics and Image Processing*, 2006

[17] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 32-36, Vol.3, August 2004.

[18] P. Dollar, V. Rabaud, G. Cottrell, and S. J. Belongie. 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)*, pages 65-72, 2005.

[19] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *Proceedings of ECCV Workshop on Spatial Coherence for Visual Motion Analysis*, pages 91-103, 2004.

[20] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of Alvey Vision Conference*, pp. 147–152, 1998.

[21] I. Laptev. On space-time interest points. *International Journal of Computer Vision (IJCV)*, 64(2-3):107-123, 2005.

[22] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol., no., pp.1-8, June 2008. Website: http://www.irisa.fr/vista/actions/.

[23] H.Wang, M. Ullah, A. Klaser, I. Laptev, C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the British Machine Vision Conference*, September 2009.

[24] P. Scovanner, S. Ali and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM International Conference on Multimedia*, September 2007. Website: http://www.cs.ucf.edu/~pscovann/

[25] M. Heikkilä, M. Pietikäinen and C. Schmid. Description of interest regions with local binary patterns. *Pattern Recognition*, vol. 42(3), pp. 425-436, March 2007.

[26] S. F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 1-8, 2007.

[27] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *Proceedings of the British Machine Vision Conference*, 2006.

[28] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[29] M. Marszalek, I. Laptev, C. Schmid. Actions in Context. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[30] R. Mattivi and L. Shao. Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor. In *Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns (CAIP)*, Munster, Germany, September 2009.