Learning Motion-Difference Features using Gaussian Restricted Boltzmann Machines for Efficient Human Action Recognition

Son N. Tran, Emmanouil Benetos, Artur d'Avila Garcez

Abstract— **Learning visual words from video frames is challenging because deciding which word to assign to each subset of frames is a difficult task. For example, two similar frames may have different meanings in describing human actions such as *starting to run* and *starting to walk*. In order to associate richer information to vector-quantization and generate visual words, several approaches have been proposed recently that use complex algorithms to extract or learn spatio-temporal features from 3-D volumes of video frames. In this paper, we propose an efficient method to use Gaussian RBM for learning motion-difference features from actions in videos. The difference between two video frames is defined by a subtraction function of one frame by another that preserves positive and negative changes, thus creating a simple spatio-temporal saliency map for an action. This subtraction function removes, by construction, the common shapes and background images that should not be relevant for action learning and recognition, and highlights the movement patterns in space, making it easier to learn the actions from such saliency maps using shallow feature learning models such as RBMs. In the experiments reported in this paper, we used a Gaussian restricted Boltzmann machine to learn the actions from saliency maps of different motion images. Despite its simplicity, the motion-difference method achieved very good performance in benchmark datasets, specifically the Weizmann dataset (98.81%) and the KTH dataset (88.89%). A comparative analysis with hand-crafted and learned features using similar classifiers indicates that motion-difference can be competitive and very efficient.**

I. Introduction

Human action recognition is considered to be a fundamental topic in computer vision research, with numerous applications in surveillance and retrieval systems [5], [13]. However, action recognition is still an open problem due to numerous associated challenges, including camera motion, occlusion, and cluttered background [26].

Typically, action recognition systems model video recordings as collections of visual words, which are estimated using hand-crafted features. Extracting features from each frame image to build codewords has been proved an efficient and useful approach. Zhang and Gong [28] used visual words generated from local shape context descriptors [1] to evaluate their proposed variant of Probabilistic Latent Semantic Analysis (PLSA) [12], namely structural PLSA. Wang and Mori [24] showed that motion descriptors [8] work perfectly on the Weizmann human action dataset [3]. Lin et al. [16] combine shape and motion descriptors to generate visual words and represent an action as a sequence of prototypes.

Son N. Tran, Emmanouil Benetos and Artur d'Avila Garcez are with the Department of Computer Science, School of Informatics, City University London, London, EC1V 0HB (email: {Son.Tran.1,Emmanouil.Benetos.1}@city.ac.uk, aag@soi.city.ac.uk)

However, spatial features alone, such as shape descriptors, are not sufficient since visual words from video should be characterized by both spatial and temporal information. Schuldt et al. [19] use local space-time features with support vector machines (SVMs) to recognize human actions in the proposed KTH dataset. Xu et al. [26] employed the incremental expectation-maximization algorithm to improve PLSA performance for action classification with the use of colour-coded space-time features. Niebles et al. [17] utilized spatio-temporal features to generate visual words and performed action modelling using PLSA.

Recently, unsupervised learning algorithms have been employed to learn spatio-temporal features from video for visual word construction. In [6], Bo et al. proposed the Space-Time Deep Belief Network, a model based on Convolutional Deep Belief Networks (CDBNs) [15] to aggregate spatial and temporal information. Le et al. [14] also applied the idea of convolution and stacking in CDBNs [15] to build a deep network that can be trained efficiently. Taylor et al. [22] employed convolutional gated Restricted Boltzmann Machines (RBMs) for learning low-level spatio-temporal features in a multi-stage architecture for action recognition.

Despite achieving good performance on benchmark datasets, per-frame visual word approaches need to extract large-scale features from every single image in the video. For example, in [28], the features extracted from a frame have dimensionality of 100×60 for 100 interest points, where the dimensionality of each point is 60. The motion descriptor [8] is a set of four channels computed from optical flow at a frame in a stabilized video sequence, with each channel having the same size as the frame. For spatio-temporal features, the data structure of video is quite complex. In unsupervised learning, it can be said that if one has good features, one only needs a simple classifier to achieve a good performance. However, learning good features from complex structures such as shape and movement in videos normally requires a complex algorithm [15], [6], [22].

In this paper, we propose a new approach to learn spatio-temporal features using a difference measure between frames in a video sequence, called motion-difference, and applying Gaussian RBMs. Motion-difference removes the common shapes and background images that should not be relevant for action learning and recognition, and highlights the movement patterns in space, making it easier to learn the actions from such saliency maps using a simple classifier. Our motion-difference is similar to motion-history [4] in that the images are constructed from sequence of frames, and therefore, be able to capture moving information of the human in a video.

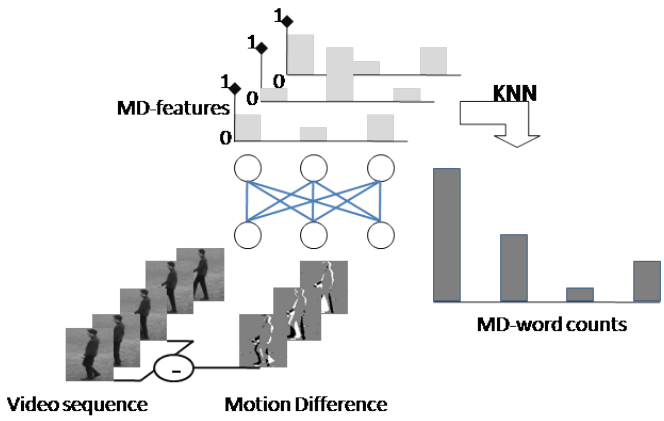


Fig. 1. Learning Motion-Difference Features for Action Recognition

However, in [4] a motion-history is an accumulative merging of every shape in consecutive frames, while in motion-difference common shapes between frames are removed. Our approach is also different from [27] in that their difference function only produces absolute values after subtraction for the purpose of edge detection. Motion-difference is able to represent both spatial information and temporal relations in a video frame by preserving negative and positive pixels in the resulting saliency map, thus indicating movement patterns and removing unnecessary common shapes and background between two frames. Despite its simplicity, the motion-difference method achieved very good performance in benchmark dataset such as the Weizmann dataset (98.81%) and the KTH dataset (88.89%). A comparative analysis with other features using simple classifiers such as Naive-Bayes and PLSA indicates that the motion-difference with Gaussian RBM is competitive with shape descriptor [28], motion descriptor [24], hand-crafted spatio-temporal features [17], [19], [26], and spatio-temporal features learned by Deep Networks [6]. In addition to good performance, our approach shows more efficicent than many other approaches [28], [24], [17], [19], [26], [6], [14], [16].

The outline of the paper is as follows. In Section II, the motion difference method is introduced and a Gaussian RBM is used to learn movement patterns from videos. Section III describes the action modelling procedure, including the bounding box extraction and classification algorithms employed. Section IV contains the experimental results and their evaluation. Section V concludes the paper and discusses directions for future work.

II. Learning Movement Features From Motion-Difference

In this section, we introduce the definition of motion-difference and the use of Gaussian Restricted Boltzmann Machines (Gaussian RBMs) to learn moving patterns from these images. The whole model is decribed in Figure 1.

A. Motion-Difference

An action video sequence consists of static images. Each of them represents a spatial distribution of the human body, however it fails to represent the temporal meaning of the action. As an example, Figure 2 shows two groups of images from video sequences of different actions. In Figures 2a and 2c, it would not be easy even for a human to know what action that the actor is going to perform from one of the images. There are alternative approaches to obtain richer information such as spatio-temporal features, however, instead of treating a video as a sequence of images, in this work we consider the video sequence as a 3-D volume image.

Recently, feature learning shows improvement in many computer vision problems [11], [15], [18]. For video image, however, learning shape is difficult to standard unsupervised models such as Restracted Boltzmann Machine [20], Deep Belief Networks [11], Deep Boltzmann Machine [18], Stacked Auto-Encoder [2]. The reason is that complex shapes such as images of people are more difficult to learn than, for example, images of characters, for which the above methods have been shown very good, in that the former have more density of foreground and different shapes share a large number of common pixels. In order to solve this problem, these models should be modified to learn small regions in images. Variant models such as convolutional approaches [15], [14] learn overlapped regions, and ShapeBM[9] learns separated image blocks. In this paper, we propose the use

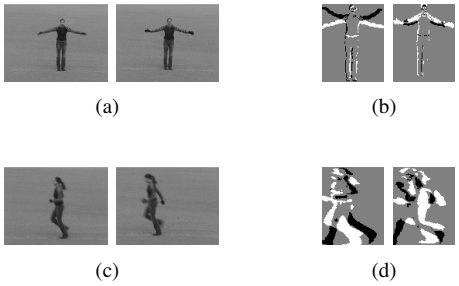


Fig. 2. Original images from four different actions. (a) Hand waving and Hand Clapping; (c) Jogging and Running (from left to right). On the right side are corresponding motion-difference images)

of motion-difference (MD),a simple method for dealing with the problem of representing temporal information and feature learning in complex video streams. A motion-difference is the subtraction of $I_{t+\kappa}$ by I_t, two images at positions $t+\kappa$ and t in a video sequence respectively, κ is frame distance.As result, in motion-difference the negative pixels show the part of an actor which only appears in the previous frame (t) while the positive pixels show the portion only in the future frame ($t+\kappa$). In other words, a motion-difference represents moving patterns as the location of negative and positive pixels. In addition, by using subtraction, the common parts between two images (including background) have been removed and therefore the motion-difference only keeps the important information which can be easily learned by feature learning models. In Figure 2, by looking at the motion-difference

one can easily understand the meaning of the movements performed by the actors.

In what follows, we apply a feature learning algorithm to motion-difference. Our expectation is that the features learned from motion-difference would capture the relations between separated movement patterns, and therefore offer a better representation.

B. Learning Movement Features With Gaussian RBM

The Gaussian RBM [25] is an RBM [20] with Gaussian visible units and binary hidden units. We denote v and h as units in visible and hidden layer respectively; W is the weight matrix of the model and σ is standard deviation of Gaussian noise for visible unit. The energy function of the Gaussian RBM is defined as:

$$\mathbb{E}(v,h) = \sum_i \frac{(v_i - a)^2}{2\sigma^2} - \frac{1}{\sigma^2}(\sum_j h_j b_j + \sum_{ij} w_{ij} v_i h_j) \quad (1)$$

Given a state of a layer, we can infer the state of units in the other layer by sampling the conditional distributions.

$$P(h_j|v) = sigmoid(\sum_i v_i w_{ij} + b_j) \quad (2)$$

$$P(v_i|h) = \mathcal{N}(\sum_j w_{ij} h_j, \sigma^2) \quad (3)$$

Where $sigmoid(x) = (1 + e^{-x})^{-1}$ is the sigmoid function and $\mathcal{N}$ is a normal distribution. Training a Gaussian RBM is similar as training an RBM by using Contrastive Divergence (CD) [10]. However, learning a and σ is difficult in CD with one step and requires more computational power. Fortunately, it is possible to normalize the data to zero mean and unit variance and train the model with $a = 0$, $\sigma^2 = 1$.

From what has been discussed in the previous subsection, learning motion-difference can be considered as learning the distribution of movement patterns $P(I_{t+\kappa} - I_t)$. Compared to learning conditional distributions $P(I_{t+\kappa}|I_t)$ to represent relation between two images in video sequence, approximate the joint distribution of motion-difference is more feasible. It is also worth noting that the features learned from Gaussian RBMs are low-level features. According to [11], the high-level features, which are as good or better than lower-level features, can be learned by stacking the shallow networks one on another. In our experiments, we apply a shallow network to learn from spatio-temporal features, and already obtain competitive results. Given the theory and recent results from deep networks, we would expect that our results could be improved further by the use of deep Gaussian RBMs. This is left as future work though

III. Action Modelling

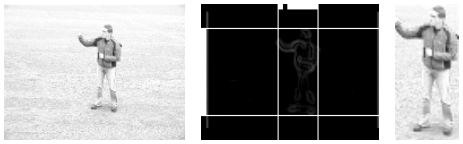
We adopt a common pipeline to model video actions [23]. At first, we stabilize the video image by extracting the bounding box surrounding the actor and apply Gaussian RBMs to learn movement features motion-difference as described in Section II-A. The features are then converted to visual words by vector quantization with KNN. After representing each video as a "bag of words" we learn efficient classifiers using Naive Bayes with smoothing and PLSA described in what follows.

A. Bounding Box Extraction

For an image with low variant background we can use gradient intensity projection to extract the bounding box surrounding an actor. In particular, to extract a bounding box from an image I we use Canny edge detection method to find the image I' that contains only the edges in I. Such a simple method is suitable with our comparative evaluation on benchmark datasets. We can use a method which is more general such as [21] for the datasets with unstable and clutter background. Let g_x and g_y denote the horizontal and vertical gradient of $G_{\mathcal{N}}(K)$, a 2-D Gaussian matrix of size K. The smoothing edge image I' is computed as root sum square of the images generated by convoluting the original image I with g_x, g_y.

$$\begin{aligned} I_x &= conv2(I, g_x) \\ I_y &= conv2(I, g_y) \\ I' &= \sqrt{(I_x * I_x + I_y * I_y)} \end{aligned} \quad (4)$$

As a result, the transformed image I' contains only edges from the original image I since all pixels in I whose neighbours have similar value to it would be converted to $\sim$ zeros in image I'. We futher reduce the noise in the edge image I' by setting all pixels with small values to zero. We project the accumulative sum of the pixels in I' onto



(a) (b) (c)

Fig. 3. Bounding box extraction process from (a): Original image to (b): edge image and projection on horizontal and vertical axes and (c): bounded image (resized)

horizontal and vertical axes and then compare the values with thresholds (τ_x, τ_y) to extract the limits of the bounding box as:

$$\begin{aligned} \mathcal{Y} &= \{y | \sum_x I'(x,y) > \tau_y\} \\ \mathcal{X} &= \{x | \sum_y I'(x,y) > \tau_x\} \end{aligned} \quad (5)$$

The bounding box is a rectangle drawn from $(min(\mathcal{X}), min(\mathcal{Y}))$ to $(max(\mathcal{X}), max(\mathcal{Y}))$

B. Classification

For classification, we employ Naive Bayes and PLSA-based classifiers because of the efficiency in their learning algorithms. In action recognition, we consider each video as a document containing visual words. More formally, we denote d as the video recording index, w as the visual word index, and z as the topic index. We also denote $n(d, w)$ as the

occurrence of visual word w in video document d, $n(z, w)$ as the occurrence of visual word w in all video documents of topic z, and $n(z)$ as the number of video documents in topic z.

Naive Bayes

By assuming that the visual words in a video document are independent given a topic, a Naive Bayes model models the conditional distribution of topic z given a video document d as:

$$P(z|d) \propto P(z) \prod_{w} P(w|z)^{n(d,w)} \tag{6}$$

in which $P(z)$ and $P(w|z)$ are computed as follows (with additive smoothing to prevent zero probabilities) to maximize the log-likelihood $\mathcal{L} = \log(\prod_{z,d} P(z, d))$:

$$\begin{aligned} P(z) &= \frac{n(d)}{\sum_{d'} n(d')} \\ P(w|z) &= \frac{n(z, w) + \alpha}{\sum_{w'} n(z, w') + \alpha \|Z\|} \end{aligned} \tag{7}$$

For classification of a new video document d_{test}, we search for the topic $\hat{z}$ such that:

$$\hat{z} = \arg\max_{z} P(z|d_{test}) \tag{8}$$

PLSA

An alternative classifier is employed which is based on probabilistic latent semantic analysis (PLSA) [12]. PLSA is a probabilistic generative model that has sucessfully been applied to action recognition (e.g. [26], [28]), and represents each video recording as a bag of words, decomposed into a probability distribution of words per 'topic' (which represents the action class) and a probability of a topic occurring in the input recording.

Given as input a word occurrence matrix $n(d, w)$, PLSA approximates it as a bivariate probability distribution $P(d, w)$ of videos and words, which is decomposed as:

$$P(d, w) = P(d)P(w|d) = P(d) \sum_{z} P(w|z)P(z|d) \tag{9}$$

where $P(d)$ is the prior probability of d (known quantity, modelled as $\sum_w n(d, w)$), $P(w|z)$ is the probability of a word given a topic, and $P(z|d)$ is the probability of a class given a document (i.e. video). The unknown parameters $P(z|d)$ and $P(w|z)$ can be estimated using the expectation-maximization (EM) algorithm [7]. In our case, $P(w|z)$ is computed during training, and only $P(z|d)$ needs to be estimated.

In specific, for the expectation step, we compute the posterior of the latent variable:

$$P(z|d, w) = \frac{P(w|z)P(z|d)}{\sum_{z'} P(w|z')P(z'|d)} \tag{10}$$

and for the maximization step, the update equation for $P(z|d)$ is:

$$P(z|d) = \frac{\sum_w n(d, w)P(z|d, w)}{\sum_{z',w'} n(d, w')P(z'|d, w')} \tag{11}$$

Equations (10)-(11) are iterated until convergence. Finally, for a test recording d_{test}, the action category is given by:

$$\hat{z} = \arg\max_{z} P(z|d_{test}) \tag{12}$$

IV. Experiments

In this section we empirically investigate the use of visual words learned from motion-difference to recognize actions in Weizmann [3] and KTH [19] datasets. In the Weizmann dataset, since the bounding box notations are already provided, we only apply the extraction method of Section III-A to the KTH dataset. After training a Gaussian RBM, the output of the hidden layer is taken as learned features for clustering visual words. We evaluate the visual words using the Naive Bayes and PLSA classifiers as described earlier. We report our results along with results from other approaches regarding to different classes of features such as shape descriptor (SD), motion descriptor (MF), hand-crafted spatio-temporal (HST), and learned spatio-temporal (LST) descriptor. In order to make a fair comparison, here we emphasise the approachs that use similar classification models such as Naive Bayes and PLSA or their invariants. For completeness, we also include the recent approaches which achieve state-of-the-art performance. Significance comparisons between the approaches is not possible since each employed different reprocessing and classification techniques. In addition, each approach adopts different method such as split or leave-one-out (l-o-o) for experimental evaluation.

A. Weizmann Dataset

The Weizmann dataset has 10 actions *"Walk", "Run", "Jump", "Gallop sideways", "Bend", "One-hand wave", "Two-hands wave", "Jump in place", "Jumping Jack", "Skip"* from 9 actors. In this experiment, we use the bounding box annotations on silhouette video images provided from the source. For evaluation, we divide the dataset into a training set consisting of actions from 4 actors and a test set with actions from the other 5 actors as similar as what has been done in [28]. Figure 4 shows the original images and the motion-difference in a video clip. In this experiment,

Fig. 4. Top row: Original video images from "jack" action; Bottom row: Motion-difference (κ=2)

we resize the motion-difference to 54×54 pixels and learn the features using a Gaussian RBM with 500 hidden units. To guarantee the generality of the performance, we run our experiment 50 times and report the average result as shown in Table I. Since the dataset is small, we did not split the data for a validation set to select a feature learning model. Instead, we fixed a Gaussian RBM with 500 units in hidden layer train it within 300 epochs. For the Naive Bayes, we

use a uniform distribution for prior and set the smoothing parameter $\alpha = 1$. With both NB and PLSA, we achieve

Method	Evaluation	Recog.rate(%)
MD + Gaussian RBM + NB	split	98.81
MD + Gaussian RBM + PLSA	split	98.77
SD + pLSA [28]	split	92.3
SD + s-pLSA [28]	split	93.00
ST + pLSA [17]	l-o-o	90.00
MF + SVM [24]	l-o-o	98.80
SD & MF + prototype tree[16]	l-o-o	100.0

TABLE I

PERFORMANCE ON WEIZMANN DATASET. THE RESULTS OF [28], [17], [24], [16] ARE COPIED FROM THE ORIGINAL PAPERS

perfect 100% accuracy in 26 out of 50 runs. The results show that good representation of data can simplify the process to model the data distribution.

B. KTH Dataset

The KTH dataset [] includes 25 actors performing 6 actions: "*boxing*", "*hand clapping*", "*hand waving*", "*jogging*", "*running*", and "*walking*" in 4 scenarios: "*Static homogeneous background*", "*Scale variations*", "*Different clothes*", "*Lighting variations*". In general, leave-one-out evaluation method may be more comprehensive, however it would be highly likely that the evaluated actor has similar features to one or more actors in training set, thus, reduce the evaluation reality. In addition, we would like to evaluate the robustness of features learned from motion-difference where some actors in the test set have different "shape" and/or colour of clothes from all other actors in training set. In this experiment, we use the split of training, validation and test sets following the procedure in [19]. The validation set is used for model selection.

We show an example of KTH video images and the corresponding difference motion images in Figure 5. It is worth noting that even though the bounding box extraction is not perfect (see the distortion in the top row images) and the background has not been removed, the different images in bottom row are able to represent interesting movement patterns.

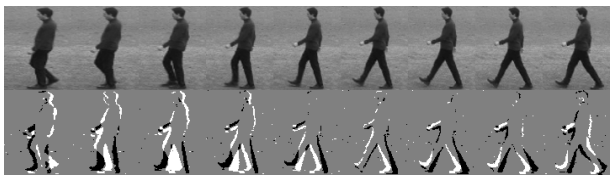


Fig. 5. (Top) Original video images from KTH dataset; (Bottom) Different motion images (κ= 2)

In this experiment we also use an uniform prior and smoothing parameter $\alpha = 1$ for Naive Bayes classifier. The PLSA is implemented following standard model [12] without the need for any hyper-parameter tuning. The Gaussian RBM is selected using validation set and average results

Approach	Evaluation	Recog.rate(%)
MD + Gaussian RBM + NB	split	85.65
MD + Gaussian RBM + pLSA	split	88.89
HST + pLSA [17]	l-o-o	83.33
MF + SVM [24]	l-o-o	83.31
HST + SVM [19]	split	71.72
HST + iEM+PLSA [26]	l-o-o	82.33
LST + SVM [6]	split	86.6
SD + S-LDA [24]	l-o-o	91.20
LST + SVM [14]	split	93.9

TABLE II

PERFORMANCE ON KTH DATASET. RESULTS OF [17], [24], [19], [26], [6], [14] ARE COPIED FROM THE ORIGINAL PAPERS

are reported along with results from [17], [24], [19], [26], [6], [14] as in Table II. Despite the simplicity, our approach is able to achieve good performance among state-of-the-art approaches.

We show the filter bases learned from motion-difference of Weizmann dataset and KTH dataset in Figure 6. Different from other approaches [26], [14] which extracted or learn local Gabor filters from video images, we learn movement patterns as visualized as pairs of back and white lines and curves.

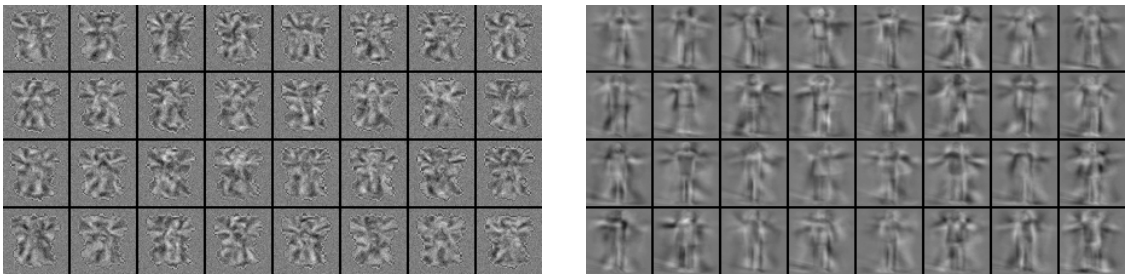
C. Model Complexity

Comparing our features with those of other methods, the features learned from motion-difference seem to have a more compact representation. For example, the dimensionality of the features learned from Weizmann and KTH datasets in our experiments have dimensionality less than or equal to 500. For other handcrafted features, the local shape contex used in [28] has the dimensionality of 6000, the dimensionality of motion descriptor [8], [24] for our 54×54 frames would be more than 10000. In [26] the spatio-temporal descriptors are so large that only two videos of each action from three actors are selected for training.

In general, learning a shallow network such as Gaussian RBM is not as computationally expensive as the convolutional variant of an RBM and other deep networks. In our experiment, it took less than 30 minutes to train a Gaussian RBM with 300 hidden units on 55000 training samples from 54×54 resized motion-difference[1]. As what has been mentioned in [14], the convolutional GRBM proposed by [22] take 2-3 days to train. The stacked convolutional ISA [14] takes 1-2 hours for 200000 input samples, each of them has the size 20×20 (spatial) and 14 (temporal).

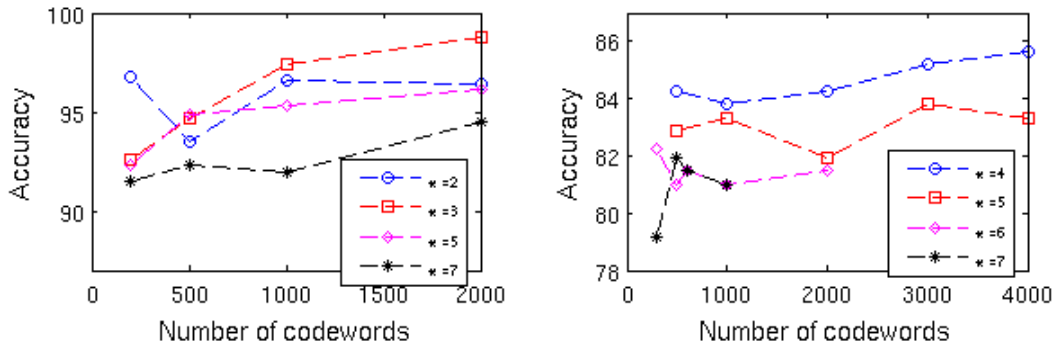
In addition, our feature extraction from a one-layer unsupervised model is very efficient. In 0.1 seconds, the Gaussian RBM is able to extract features from 1000 frames each has 54 pixels (totalling 2916000 pixels) . At the same amount of time, stacked convolutional ISA with GPU sport can only extract features from dense samples of a single 360×288 frame (totalling103680 pixels)

[1]Our system is implemented in MATLAB without using parallel computing support



(a) Weizmman dataset (b) KTH dataset

Fig. 6. Visualization of 24 filter bases from GRBMs trained on motion-difference of (a) Weizmann and (b) KTH datasets



(a) Weizmann dataset (b) KTH dataset

Fig. 7. Classification accuracy with Naive Bayes versus codebook size and frame distance of motion-difference

In what concerns the use of different codebook sizes, according to the evaluations among difference approach [23], a codebook size of 4000 is applicable to wide range of datasets. In our experiment as shown in Figure 7, for Weizmann dataset, the system achieved decent performance with number of codewords less than or equal to 2000. For KTH dataset, the best results has been achieved from the codebook size of 4000 but with 1000 codewords the system can already produce reasonable levels of accuracy. Finally, in our system, we use very efficient classifiers. With Naive Bayes and standard PLSA, a newly arrived video document can be categorized in matters of second.

V. Conclusions

We proposed a new approach to learn spatio-temporal features using a difference measure between frames in a video sequence, called motion-difference, and applying Gaussian RBMs. Motion-difference is able to represent both spatial information and temporal relations in a video frame by preserving negative and positive pixels in the resulting saliency map, thus indicating movement patterns and removing unnecessary common shapes and background between two frames. To our best knowledge, this is the first work on frame-based feature learning for action recognition. Despite its simplicity, the motion-difference method achieved very good performance in the Weizmann dataset and the KTH dataset.

In our approach, the motion-difference is sparse and then easier to learn using standard model such as RBM than denser shape images. Its efficiency shows considerable promise to application to larger datasets and the use of deeper models in learning movement features.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, April 2002.

[2] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, pages 153–160, 2006.

[3] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402, 2005.

[4] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267, March 2001.

[5] G. J. Burghouts, H. Bouma, R. den Hollander, B. van den Broek, and K. Schutte. Recognition of 48 human behaviors from video. In *In Proc. of International Symposium on Optronics in Defence and Security (OPTRO)*, 2012.

[6] Bo Chen, Jo-Anne Ting, Benjamin Marlin, and Nando de Freitas. Deep learning of invariant spatio-temporal features from video. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning,*, 2010.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[8] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 726–, Washington, DC, USA, 2003. IEEE Computer Society.

[9] S. M. Ali Eslami, Nicolas Heess, and John M. Winn. The shape boltzmann machine: A strong model of object shape. In *CVPR*, pages 406–413. IEEE, 2012.

[10] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, August 2002.

[11] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):15271554, July 2006.

[12] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.

[13] 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*, 35(1):221–231, 2013.

[14] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368, 2011.

[15] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 609–616, New York, NY, USA, 2009. ACM.

[16] Zhe Lin, Zhuolin Jiang, and Larry S. Davis. Recognizing actions by shape-motion prototype trees. In *ICCV*, pages 444–451. IEEE, 2009.

[17] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vision*, 79(3):299–318, 2008.

[18] Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep boltzmann machines. *Journal of Machine Learning Research - Proceedings Track*, 5:448–455, 2009.

[19] Christian Schuldt, Ivan Laptev, and Barbara Caputo. 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, 2004. IEEE Computer Society.

[20] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In *In Rumelhart, D. E. and McClelland, J. L., editors, Parallel Distributed Processing: Volume 1: Foundations*, pages 194–281. MIT Press, Cambridge, 1986.

[21] Chris Stauffer and W. Eric L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, pages 2246–2252, 1999.

[22] GrahamW. Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision ECCV 2010*, volume 6316 of *Lecture Notes in Computer Science*, pages 140–153. Springer Berlin Heidelberg, 2010.

[23] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the British Machine Vision Conference*, pages 124.1–124.11. BMVA Press, 2009. doi:10.5244/C.23.124.

[24] Yang Wang and Greg Mori. Human action recognition by semilatent topic models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10):1762–1774, October 2009.

[25] Max Welling, Michal Rosen-Zvi, and Geoffrey E. Hinton. Exponential family harmoniums with an application to information retrieval. In *NIPS*, 2004.

[26] Jie Xu, Getian Ye, Yang Wang, Gunawan Herman, Bang Zhang, and Jun Yang. Incremental EM for probabilistic latent semantic analysis on human action recognition. In *Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 55–60, 2009.

[27] Ming Yang, Fengjun Lv, Wei Xu, Kai Yu, and Yihong Gong. Human action detection by boosting efficient motion features. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 522–529, 2009.

[28] Jianguo Zhang and Shaogang Gong. Action categorization by structural probabilistic latent semantic analysis. *Computer Vision and Image Understanding*, 114(8):857–864, August 2010.