Jurnal Muara
Sains, Teknologi, Kedokteran, dan Ilmu Kesehatan
Vol. 1, No. 1, April 2017: hlm 218-226

ISSN 2579-6402 (Versi Cetak)
ISSN-L 2579-6410 (Versi Elektronik)

BAG OF WORDS APPROACH AND DOCUMENT-TOPIC MODELING FOR HUMAN ACTIVITY RECOGNITION FROM VIDEOS

Janson Hendryli
Informatics Engineering, Faculty of Information Technology, Tarumanagara University
jansonh@fti.untar.ac.id

ABSTRACT

Human activity recognition from videos have many useful real world applications, ranging from multimedia, entertainment, and security. In this paper, an approach inspired by a popular text document, namely the bag of words and document topic modeling, is explored. The latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF) are used to model the latent topic distribution in videos. Finally, the discovered distribution can be used to transformed the bag of words representation in order to categorize the video into ten daily human activities. The classification is done by feeding the transformed term-frequency of the visual words to the logistic regression and SVM model. The NMF achieved higher F1-score than the LDA when both SVM and logistic regression is used as the classifier.

Keywords: human activity recognition, bag of words, document topic modeling

1. INTRODUCTION

The successful of human activity recognition from videos have many applications in the real world settings. Some of the most prominent applications are for automatic video surveillance, human-computer interaction, visual multimedia, and video games. There are many challenging aspects of recognizing and categorizing human activities from a video, such as the sheer volume of possible human activities, numerous variations and complexity of a movement that describe an activity, and also the difficulties of detecting various human body configuration (Tran, 2008).

Previous research had focused on the stochastic sequence of actions in a video to model the human behavior (Robertson, 2006). The stochastic sequences are usually modeled using hidden Markov model (Brand, 2000) or Bayesian network (Buxton, 2003; Town, 2004). To be able to detect activity, a model has to detect important cues from a video which can differentiate one activity from another. These important cues, or features, have been the subject of many studies. Some of the popular ones are local motion descriptors such as SIFT (Lowe, 2004) and SURF (Bay, 2008). Space-time interest point detector such as in Laptev (2005) expands the detector to the video domain and extracts the histogram of oriented gradient (HoG) and the histogram of optic flow (HoF) as the local motion descriptors.

This paper explores an approach to activity recognition from a video using techniques widely used on text document: the bag of words and document topic modeling. Malgireddy (2013) argued that this approach holds a major advantage over existing approaches in its ability to recognize activities with complex structures. This approach represents a video as a collection of feature points that is detected by a space-time interest point detector. The feature points are considered as if they are words in a document (with video as a document). Each feature point, or sometimes are also called visual words, will be described by HoG and HoF around it, and then clustered to form a codebook of vocabulary. After that, a document topic modeling approach is used to model latent topic distribution from the video corpus and the categorization of the human activities from the video can be done by the classifier.

2. BAG OF WORDS APPROACH AND DOCUMENT TOPIC MODELING

Bag of words model is frequently used in information retrieval and text processing model where a text document is represented as a bag of its words without regarding the grammar or the order

of the words. Recent papers (Niebles, 2008; Yang, 2007) explore the bag of words approach to computer vision problems, namely detecting objects on images and for human activity recognition. The first step of generating the bag of words for activity recognition is to detect important features which represent the activity in the best way possible. These features are then clustered into a codebook of vocabulary and represented as visual words in a video.

In this paper, the Harris3D corner detector (Laptev, 2005) is used to detect interest points from the videos. The Harris3D interest point detector is an extension of Harris corner detector that is widely used in image processing. Interesting events on a video can be represented by the moving corners of the objects in the video. The Harris3D compute local and spatio-temporal interest point based on where a pixel has significant changes in directions (H. Wang, 2009). The interest points can be detected by the high ratio of the eigenvalues λ_2/λ_1 in equation 1 below:

$$\begin{aligned} H^{sp} &= \det(\mu^{sp}) - k \cdot \operatorname{trace}(\mu^{sp}) \\ &= \lambda_1\lambda_2 - k \cdot (\lambda_1 + \lambda_2)^2 \end{aligned} \tag{1}$$

where μ^{sp} is the convolution of the Gaussian-smoothed image and its second-order Gaussian derivatives, and k is a user-defined constant. Figure 1 and 2 below demonstrates the detected interest points on videos of a person answering phone and drinking water from our dataset. The interest points are circled in yellow, which show a moving corner in consecutive frames.

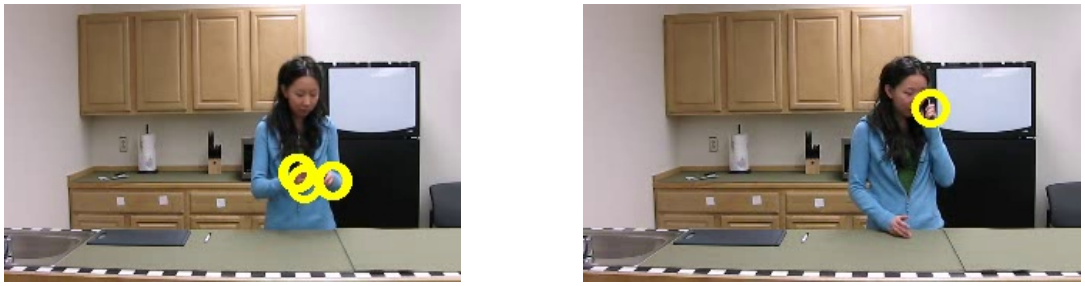

Figure 1. Example of interest points detected on two frames of a video of a person answering phone

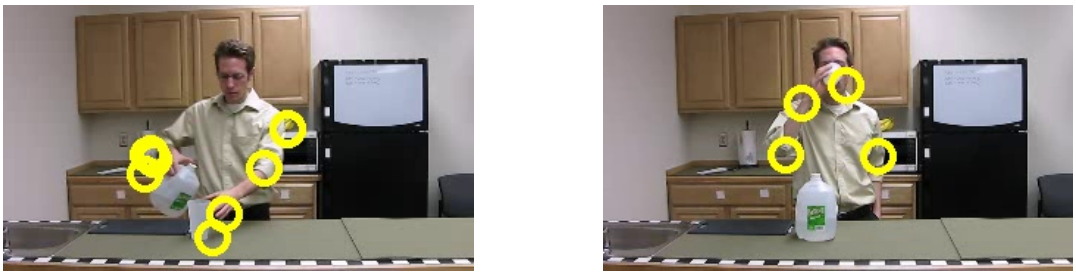

Figure 2. Example of interest points detected on two frames of a video of a person pouring and drinking water from a glass

For each interest points on the video, the histogram of oriented gradients (HoG) and the histogram of optic flows (HoF) are extracted as the descriptors around the detected points. The joint HoG and HoF of every interest points are then normalized and clustered using K-means

ISSN 2579-6402 (Versi Cetak)
ISSN-L 2579-6410 (Versi Elektronik)

algorithm. The cluster of each interest points, which described by the concatenation of HoG and HoF, constitutes the vocabulary of a video. For example, a video V_i can be represented as a vector of visual words $(C_1, C_2, \ldots, C_K)$, where K is the number of detected interest points and $C_k,\ k \in K$ is the cluster of interest point k.

Visualization in Figure 3 uses the 2-dimensional linear PCA and non-linear T-SNE model (Maaten, 2008) with a perplexity of 2 learned from the term-frequency of the corpus to show the difficulty of categorizing 10 daily human activities only from the bag of words. From the PCA result, there are some activities clustered in a shorter distance within each other, such as drinking water, using silverware, and peeling a banana. Meanwhile, the eating snack activity is completely separated from the other. In contrast, the t-SNE cannot clearly separate the activities, which could give an insight that the data are not embedded in non-linear space.

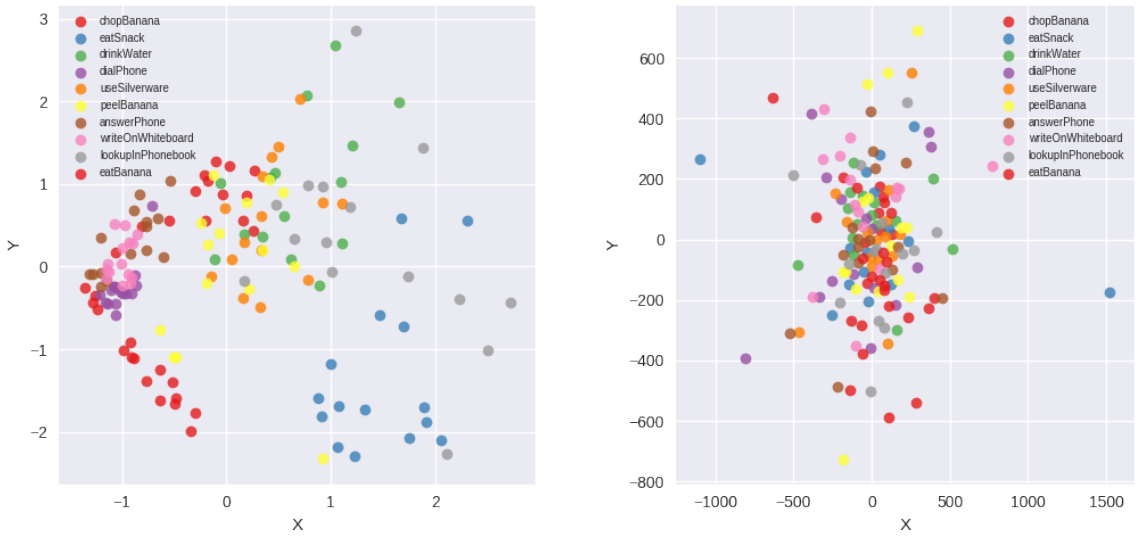

Figure 3. The visualization of PCA (left) and t-SNE (right) model for the term-frequency of the video corpus

To correctly classify the activity from the videos, the document topic modeling approach is explored in this paper. The models are the latent Dirichlet allocation (LDA) and the non-negative matrix factorization (NMF). The LDA was introduced by Blei (2003) as a generative probabilistic model for discovering the topics of documents in a corpus. The basic idea of LDA is that documents are represented as random mixtures over latent topics; and each topic, in turn, is characterized by a multinomial distribution over words (Blei, 2003).

The generative process of LDA can be seen as in Figure 4. Fitting the term-frequency to the LDA model gives the model parameter α and β that maximizes the log likelihood $\ell(\alpha,\beta) = \sum_{d=1}^{M} \log P(w_d|\alpha,\beta)$ where $P(w_d|\alpha,\beta)$ constitutes the marginal distribution of a document (Blei, 2003). From the learned LDA model, the document topic distribution for a video can be used as the input to the discriminative classifier, such as the logistic regression model or the SVM to recognize the human activities on a video.

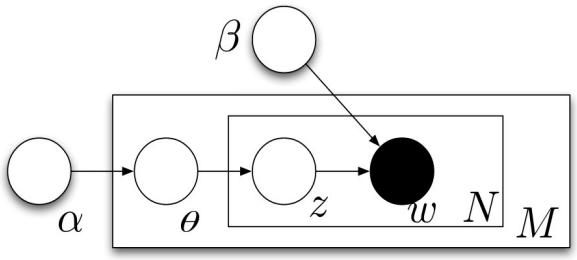

Figure 4. The graphical model of latent Dirichlet allocation (Y. Wang, 2009).

Another approach also explored in this paper is the non-negative matrix factorization (NMF), which is also frequently used for dimensionality reduction, music transcription (Smaragdis, 2003), and document clustering (Xu, 2003). Given a non-negative matrix X, the NMF finds two non-negative matrices, W and H, such that the product of those two matrices approximates X (Eq. 2).

$$X = W\ H \tag{2}$$

The NMF basically discover latent structure in the data that approximate the linear combination of the columns of W for each data vector $x \in X$, weighted by the components of $h \in H$ (Lee, 2001) by minimizing the objective function:

$$J = \frac{1}{2}\left\| X - WH^{T} \right\| \tag{3}$$

The fitted NMF is then used to transform the term-frequency vector for each video and fed as the input to the classification model.

3. EXPERIMENTS AND RESULTS

The experiment uses University of Rochester activities of daily living dataset (Messing, 2009) which consists of 150 videos of 10 different human daily activities (15 videos for each activity), such as answering phone, chopping banana, dialing a phone, drinking water, eating banana, eating snack, looking up in a phonebook, peeling banana, using silverware, and writing on a whiteboard. For the computational speed purpose, the videos are resized from a high resolution of 1280 x 760 pixels at 30 frames per second (fps) to 320 x 180 pixels (no fps changes). The visual words and codebook of vocabulary are generated as explained in the previous section with the number of activity class sets to 10. The dataset is then split into training and testing set in a stratified way with 25% data is held for the test set.

The implementations of K-means clustering, LDA, and NMF are from the scikit-learn library. In a similar way as Malgireddy (2013), we set the number of topics for the K-means to 1000. After that, the LDA model is learned in batch for 100 iterations; the regularization constant and mixing parameter of NMF are set to 0.1 and 0.5 respectively. After transforming the term-frequency vector of the video corpus, the classification result of logistic regression model and SVM are compared. The logistic regression model is learned using regularization strength of 1.0e+005, Newton conjugate gradient method for the multinomial loss optimization. Meanwhile, the SVM uses a linear kernel and a default 1.0 penalty parameter for the error term.

ISSN 2579-6402 (Versi Cetak)
ISSN-L 2579-6410 (Versi Elektronik)

The precision, recall, and F1-score will be used as the metric to evaluate the result. Table 1 and 2 below describes the precision, recall, and F1-score for the logistic regression and SVM classifier using both LDA and NMF for the document topic modeling. Generally, the NMF gives a better result than the LDA for both classifiers, with F1-score of 71% on logistic regression and 80% on SVM. Meanwhile, LDA suffers a significant drop of F1-score when the logistic regression model (63%) is changed into SVM (49%). Also, inspecting the precision and recall of both LDA and NMF gives insight that the precision of NMF is higher than the recall. In contrast, the precision value of LDA is lower than the recall.

Table 1. Precision, Recall, and F1-score of Logistic Regression Model

Activity	LDA			NMF		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
answerPhone	0.50	0.50	0.50	0.25	0.25	0.25
chopBanana	0.75	0.75	0.75	0.75	0.75	0.75
dialPhone	0.75	0.75	0.75	0.67	1.00	0.80
drinkWater	0.80	1.00	0.89	1.00	0.75	0.86
eatBanana	1.00	0.67	0.80	1.00	1.00	1.00
eatSnack	0.75	0.75	0.75	0.67	1.00	0.80
lookupInPhonebook	0.50	0.75	0.60	1.00	0.75	0.86
peelBanana	0.00	0.00	0.00	0.50	0.50	0.50
useSilverware	0.20	0.33	0.25	0.67	0.67	0.67
writeOnWhiteboard	1.00	1.00	1.00	1.00	0.50	0.67
Average	**0.63**	**0.66**	**0.63**	**0.75**	**0.71**	**0.71**

Moreover, Figure 5-8 show the confusion matrix from the experiments. The LDA misclassifies all eatBanana videos and predicted it as chopBanana, dialPhone, and answerPhone. The majority of the answerPhone videos are also misclassified as useSilverware. The misclassification of chopBanana plays a huge role in the drop of F1-score from logistic regression model to the SVM. At the same time, the NMF + logistic regression also has difficulties in classifying the chopBanana, while the NMF + SVM classifies it correctly. Finally, the NMF + SVM model with the highest score find it challenging to classify the drinkWater activity, mistook it to the eatSnack and eatBanana activity.

Table 2. Precision, Recall, and F1-score of SVM Model

Activity	LDA			NMF		
	Precision	Recall	F1-score	Precision	Recall	F1-score
answerPhone	0.00	0.00	0.00	0.80	1.00	0.89
chopBanana	0.33	0.50	0.40	1.00	0.50	0.67
dialPhone	0.33	0.25	0.29	0.67	1.00	0.80
drinkWater	1.00	1.00	1.00	1.00	0.75	0.86
eatBanana	0.60	1.00	0.75	1.00	0.33	0.50
eatSnack	0.75	0.75	0.75	0.80	1.00	0.89
lookupInPhonebook	0.40	0.50	0.44	1.00	1.00	1.00
peelBanana	0.00	0.00	0.00	0.50	0.50	0.50
useSilverware	0.17	0.33	0.22	0.75	1.00	0.86
writeOnWhiteboard	1.00	1.00	1.00	1.00	1.00	1.00
Average	**0.46**	**0.53**	**0.49**	**0.85**	**0.82**	**0.80**

Figure 5. Confusion matrix of LDA with logistic regression classifier

ISSN 2579-6402 (Versi Cetak)
ISSN-L 2579-6410 (Versi Elektronik)

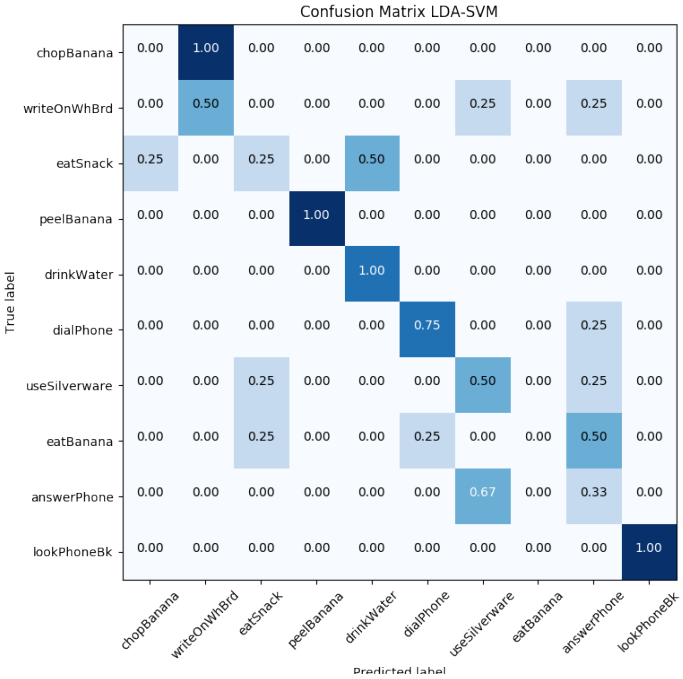

Figure 6. Confusion matrix of LDA with SVM classifier

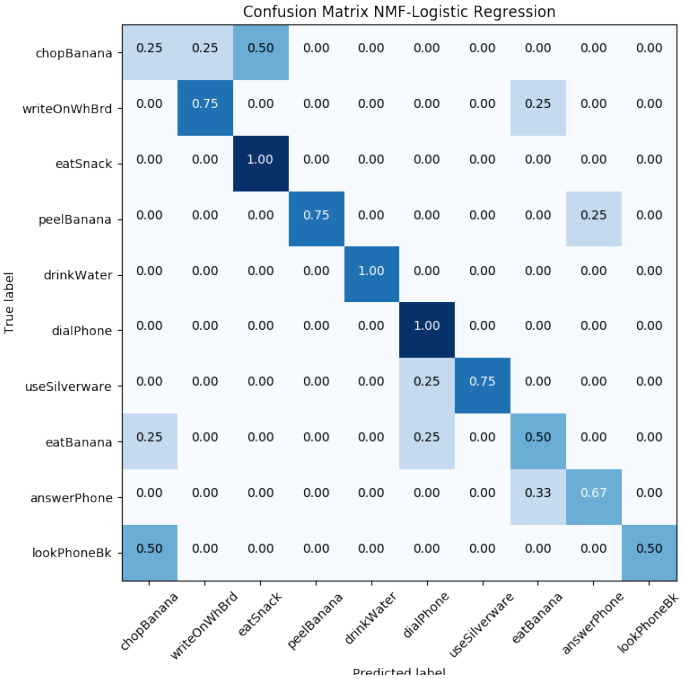

Figure 7. Confusion matrix of NMF with logistic regression classifier

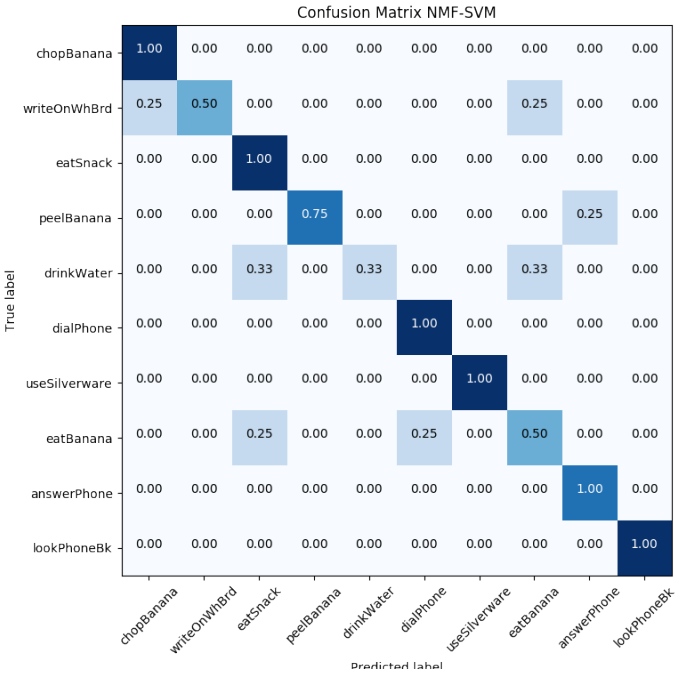

Figure 8. Confusion matrix of NMF with SVM classifier

4. CONCLUSIONS

This paper explored the bag of words approach and document topic modeling for recognizing human activity in a video. Two models for document topic modeling: the latent Dirichlet allocation model (LDA) and the non-negative matrix factorization (NMF), are used to discover latent topics from the video corpus. The transformed term-frequency from the learned LDA and NMF model is then used for multi-class classification, categorizing each video into 10 daily human activities. From two classification model: the logistic regression and SVM, it is found that the NMF with SVM gives convincingly better result than the LDA with both classifiers. The SVM significantly improves classification accuracy of NMF, but the same improvement does not apply to LDA.

The methods presented in this paper are still assuming that a video is a collection of visual words, disregarding the sequence of its occurrence, i.e. the temporal aspects of a video. Although the NMF with SVM classifier gives a good result for this particular dataset, in the future works, it is required to study the spatio-temporal representation of a video to get a better generalization of a human activity recognition system.

REFERENCES

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Brand, M., & Kettnaker, V. (2000). Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 844-851.

Buxton, H. (2003). Learning and understanding dynamic scene activity: A review. *Image and vision computing*, *21*(1), 125-136.

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

ISSN 2579-6402 (Versi Cetak)
ISSN-L 2579-6410 (Versi Elektronik)

Lee, D. D., & Seung, H. S. (2001). "Algorithms for non-negative matrix factorization". Advances in Neural Information Processing Systems, Vancouver, Canada, 3-8 December 2001, 556-562.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Key- points. *International Journal of Computer Vision*, *60*(2), 91–110.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579-2605.

Malgireddy, M. R., Nwogu, I., & Govindaraju, V. (2013). Language-motivated approaches to action recognition. *Journal of Machine Learning Research*, *14*(1), 2189-2212.

Messing, R., Pal, C., & Kautz, H. (2009). "Activity recognition using the velocity histories of tracked keypoints". International Conference on Computer Vision, Kyoto, Japan, 29 September – 2 October 2009, 104-111.

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, *79*(3), 299-318.

Robertson, N., & Reid, I. (2006). A general method for human activity recognition in video. *Computer Vision and Image Understanding*, *104*(2), 232-248.

Smaragdis, P., & Brown, J. C. (2003). "Non-negative matrix factorization for polyphonic music transcription". Workshop on Applications of Signal Processing to Audio and Acoustics, New York, United States, 19-22 October 2003, 177-180.

Town, C. (2004). "Ontology-driven Bayesian networks for dynamic scene understanding". Computer Vision and Pattern Recognition Workshop, Washington DC, United States, 27 June – 2 July 2004, 116-116.

Tran, D., & Sorokin, A. (2008). "Human activity recognition with metric learning". European Conference on Computer Vision, Marseille, France, 12-18 October 2008, 548-561.

Wang, H., Ullah, M. M., Klaser, A., Laptev, I., & Schmid, C. (2009). "Evaluation of local spatio-temporal features for action recognition". British Machine Vision Conference, London, UK, 7-10 September 2009, 124-1.

Wang, Y., Sabzmeydani, P., & Mori, G. (2007). Semi-latent dirichlet allocation: A hierarchical model for human action recognition. *Human Motion—Understanding, Modeling, Capture, and Animation*, 240-254.

Wang, Y., & Mori, G. (2009). Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(10), 1762-1774.

Xu, W., Liu, X., & Gong, Y. (2003). "Document clustering based on non-negative matrix factorization". Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 28 July – 1 Agustus 2003, 267-273.

Yang, J., Jiang, Y. G., Hauptmann, A. G., & Ngo, C. W. (2007). "Evaluating bag-of-visual-words representations in scene classification". Workshop on Multimedia Information Retrieval, Augsburg, Germany, 28-29 September 2007, 197-206.