Paper Review of *Multi-View Super Vector for Action Recognition[1]*

Justin Baraboo

**Abstract**— There are many different features used for image and video analysis and action recognition. This paper looks into creating a fusion scheme, Multi-View Super Vector, to model the shared and private information between two features, which differs from other fusion schemes. Multi-View Super Vector is born from a lineage of Cannonical Correlation Analysis which is briefly described and illustrated. This fusion scheme outperforms other fusion schemes aggregated by VLAD or Fisher Vector as well as the state of the art action recognition. It’s tested on two different video data sets HMDB51 and UCF101.

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

—————————— ◆ ——————————

# 1 Introduction

Image and video analysis is a feature rich environment, growing larger as new features extend or measure new descriptors. This can be easily seen through the lineage of descriptors covered in class: color histogram as a color descriptor, histogram of oriented gradients as a texture descriptor, and SIFT points as a texture-keypoint descriptor.

The general narrative of extending features has been to craft a better feature, outperforming the previous, and use that as a descriptor; however, features both learned in class and covered in topic papers (such as ORB, BRISK, etc) are not totally eclipsed by the ‘best’ feature: features tend to capture different information about the system and can become more descriptive and information rich when that part of the image is the most dominant. It is evident that for a model to be robust and more accurate that a fusion of descriptors, especially those that capture different information, is needed.

This is especially pertinent for video data and action recognition. The feature environment is further extended to include static image analysis but also dynamic feature descriptors such as histogram of optical flow, motion boundary histograms, and spatial temporal interest points (STIP). Here, static features and dynamic features will measure very different aspects and provide different information; hence, fusing these types of features is ideal.

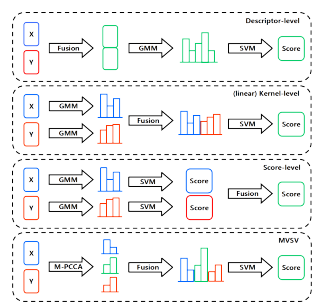


Figure 1. Descriptor Fusion Schemes

Local spatio-temporal features have great success in action recognition. Local sampling and projections through dense projection and ω-trajectory can be used in tandem with feature fusion to provide more rich features to fuse.

Multi-View Super Vector is mathematically crafted from Canonical Correlation Analysis. By reforming the problem in a probabilistic view, a latent vector capturing the shared information of the features can be found, and then stitching together many different latent vectors, solving CCA locally, will give rise to a global latent vector. To complete the Multi-View Super Vector, the gradients of the features which encode the private information are computed.

# 2 Descriptor Fusion

While the power of descriptor fusion is apparent, the implementation of how to fuse descriptors is murky. Different schemes of fusion are available and researched; four different schemes were investigated by this paper (l). Descriptor level fusion is the simplest of the fusion schemes where features are fused and then aggregated and then label guided boundary decisions are used before generated a score or prediction. We performed a similar fusion in our first homework assignment with HoG and color histogram. Linear kernel fusion first aggregates the features using a Gaussian mixture model and then performs a fusion where it then decides boundaries and makes a scoring/prediction. Score level fusion computes a score for each of the features separately and then appropriately fuses the scores together via their confidence scores. Finally, multi view super vector, determines the shared information, through a latent vector Z, and their private information, through gradient descriptors, fuses this information and then performs support vector machine to develop a score/prediction.

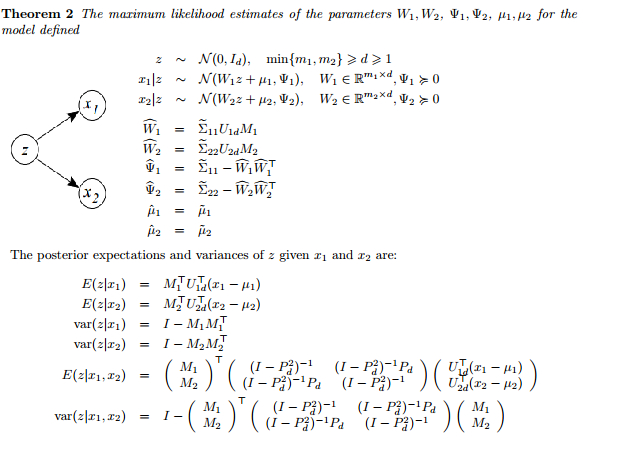


Figure 2. Mathematical Formulation of PCCA from Bach paper. Here Z is the latent vector which generates the features X1 and X2

The variety of fusion schemes comes from how independent or correlated the feature descriptors are. If the descriptors are correlated, then descriptor level fusion works well as correlation between the descriptors is taken into account by the fusion method. However, if they are independent, kernel level fusion is better as bias in one descriptor might also affect another if they were fused before aggregation. In practice, features tend to be not fully independent but not fully correlated with one another. This creates problems with both descriptor and kernel level methods; multi view super vector was created to remedy this problem, allowing it to be more robust with many different pairs of features.

.

# 3 Canonoical Correlation Ensemble

## 3.1 Canonoical Correlation Analysis

As demonstrated in this document, the numbering of sec

Canonical correlation analysis (CCA) is the basic unit of fusion for Multi-View Super Vector’s encoding of shared information through a latent vector. CCA is a well-defined and utilized statistical method, being used to find projections to maximize correlation. CCA works on two different features X and Y with dimensions x and y, respectively, where it tries to find a linear transformation to maximize the correlation between X and Y. Mathematically, this is done by maximizing the correlation, defined by the covariance of X and Y divided by the standard deviations of both X and Y, of canonical pairs.

## 3.2 Probabilistic Canonical Correlation Analysis

This technique is further framed into a probabilistic approach termed: probabilistic canonical correlation analysis. This was explored by [2] and further extended to re explain the problem in terms of having a latent vector Z that will generate X and Y and that X and Y are independent once Z is given. This idea forms the backbone of Multi-View Super Vector and their paper could be reviewed in its own right. A brief outline is given below.

Given 2 features, X and Y, they show that if one has a feature Z, with several constraints, then canonical correlation directions (or projection) can be found through a maximum likelihood calculation with a latent vector z. This is summed up by a theorem presented in their paper (figure 2). The most important part of this is the latent vector as the feature descriptors can be measured and the latent vector finds the information between them that is the space they share.

## 3.3 Multiple Probabilistic Canonical Correlation Analysis

PCCA is not enough though as it is still a linear projection. To allow for nonlinearity, a mixture model of PCCA local models are stitched together. So PCCA is applied locally, and the latent vectors for each local area are concatenated to give a global latent vector Z. How do we decide where to split and project? The entire point of defining things probabilistically aims to reformat the problem so that applying an expectation maximization solution is easier.

The expectation maximization algorithm of learning the solution works as following. K means is used to find the mean and then the corresponding variance of each local cluster. In the expectation state, the responsibility γi,k is esimtated, meaning the responsibility that the kth submodel generates the ith sample. The latent vectors for each local model is computed from above from PCCA. In the Maximization state, the means, variances, weights, and projection matrices are updated based upon the assignment to maximize the complete data log likhilood. This process is then repeated until convergence.

# 4 Multi-View Super Vector

The Multi-View Super Vector is composed of the latent vector from M-PCCA and gradients from features; the latent vector encodes information shared information between the features and the gradients encode the private information for each of the features. The concatenation of the local latent vectors to create a global latent vector uses a sum pooling scheme.

The construction of the gradients of the features uses the feature space of each descriptor using the local means, variances, and projection matrices (these feature ensembles are called λx and λy).

This construction is defined by taking the derivative of the complete log-likelihood of M-PCCA with respect to {µ k , Ψk} to obtain the gradient vector G, doing so for each feature. In explicit mathematical terms it is written as:

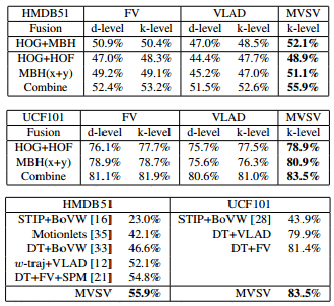


Figure 3. Experimental Results of MVSV. First two tables correspont to FV vs VLAD schemes of different features for different fusion schemses: d- descriptor, k-kernel. The third table shows how MVSV compares to state of the art methods.

∂E(L) /∂µk x = −2 X i γi,kxi,k/σk x

∂E(L)/∂σk x = 2(wkσ˜ k x − diag( X i γi,kxi,kx > i,k))/σk x

Where

wk = Σγi,k, σ k x = p diag(Ψk x )

σ˜ k x = σ k x − diag(Wk x Σ k zWk x > )

xi,k = xi − µ k x − Wk x z˜i,k

Thus the representation can be written as:

MVSV = {Z, Gx, Gy}.

These features are further L2 normalized and has a dimension of K(n+m+d) (K local models, n/m dimension of features, d dimension of latent vector).

# 5 Reduction to VLAD and Fisher Vector

Multi-View Super Vector is also special in that VLAD and to an extent Fisher Vector are specific cases of it. In the case of VLAD, we can reduce to it by finding k-means and variances locally to learn local means and variances. For each local model, the projection matrices are found using CCA and the weighting scheme and the weighting coefficient γt,k is set to 1, if the nearest neighbor of xt is µk, and 0 otherwise; thus the local latent vector becomes like the aggregator centroid similar to VLAD.

For Fisher Vector, the reduction is a little more nuanced. Fisher Vector uses means and variances to aggregate and assign points. Here Multi-View Super Vector is probabilistic counter, using a latent vector, similar to mean, and gradients which describes the direction in which parameters should be stretched to best fit the data, instead of variances.

# 6 Experiment And Results

The data sets evaluated for this paper were the HDMB51 video data set and the UCF101 video data set. Each of these data sets contain various action content and is separated accordingly. The HDMB51 data set contains 6,766 videos and 51 action categories while the UCF101 13,320 videos and 101 action classes. The videos themselves are not all uniform length. The HDMB51 data set was split into three groups where each group had a 70-30 test-train split. The UCF101 database was split into 25 groups

Dense trajectories were used to extract local features and the key features extracted were the motion boundary histograms, histogram of optical flow, and histogram of oriented gradients. These features underwent principal component analysis to half their dimensions, were whitened, and then l2 normalized, that is each component in a feature vector was divided by the l2-norm of the vector. 256,000 features were sampled from their training data to form the model.

Multi-View Super Vector has to have a few parameters defined before being able to perform feature fusion. The latent space dimension needs to be defined and the number of components. It was empirically determined that a 45 dimension latent vector and 256 components worked the best.

The Multi-View Super Vector scheme was tested against VLAD and Fisher Vector aggregation fusing the descriptors at different levels as well as modern action recognition techniques (figure 3). Multi-View Super Vector outperforms VLAD and Fisher Vector regardless of aggregation scheme using the same descriptors. What really tests its mettle is its ability to outperform the modern action recognition techniques, even if it is just narrowly.

# 7 Conclusion

Multi-View Super Vector is able to fuse together features in a meaningful way by modeling the shared information between two features and their private information in an appropriate way. This is shown in both its derivation through MPCCA and gradient based design and in its testing where it outperforms the other aggregation schemes and state of the art descriptors. The major drawback from this that it only works for 2 features. Anything more than that gets even more mathematically complicated as features will have pairwise shared information, total shared information, and private information, growing exponentially in complexity.

Encoding and fusing private and shared information almost begs if mutual information or information entropy could be utilized to create fusing hierarchy. Such measures have been used to dimension reduction to find a subset of features that contains fewer features but perserves information is well researched, but the idea of using it to aggregate features isn’t well known.

**References**

1. “Multi-View Super Vector for Action Recognition” Zhuowei Cai, Limin Wang, Xiaojiang Peng, Yu Qiao.   
   IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014.
2. “A Probabilistic Interpretation of Canonical Correlation Analysis” F. R. Bach and M. I. Jordan. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.