

# LEARNING CROSS-MODAL APPEARANCE MODELS WITH APPLICATION TO TRACKING

John W. Fisher III

Massachusetts Institute of Technology  
Artificial Intelligence Laboratory  
Cambridge, MA

Trevor Darrell

Massachusetts Institute of Technology  
Artificial Intelligence Laboratory  
Cambridge, MA

## Abstract

*Objects of interest are rarely silent or invisible. Analysis of multi-modal signal generation from a single object represents a rich and challenging area for smart sensor arrays. We consider the problem of simultaneously learning and audio and visual appearance model of a moving subject. We present a method which successfully learns such a model without benefit of hand initialization using only the associated audio signal to "decide" which object to model and track. We are interested in particular in modeling joint audio and video variation, such as produced by a speaking face. We present an algorithm and experimental results of a human speaker moving in a scene.*

## 1. INTRODUCTION

Objects of interest are rarely silent or invisible. Efforts to model and perceive them will therefore be more effective if they can consider the signals an object generates across multiple modalities. We are interested in particular in modeling joint audio and video variation, such as produced by a speaking face. For many tasks, including detection, localization, and identification, the use of information from both modalities can make processing both more accurate and robust. In this light we consider "smart cameras" to be "smart sensors" which can adapt to their users and their environment in order to integrate multiple information sources in an intelligent manner.

Consider an image sequence in which changes are attributable to several objects which undergo rigid body and/or iconic motion. We wish to explain observed changes in the scene via a *cross-modal* appearance model of finite support (e.g. a set of appearance bases) which we will learn from the sequence itself. We are liable to learn a model of *any* of the objects in the scene using visual appearance alone in the absence of additional information and explicit initialization. Suppose we have an associated signal, specifically an audio signal, which we know to be associated with one object in the scene. We discuss a statistical learning approach by which one can use the associated audio signal as an indirect pointer to the object of interest in the scene and by which one can develop a combined audio-visual appearance model for that object. Ultimately, one might like to decompose a complicated scene into independent multi-modal data streams such that they

---

This work supported in part by MIT Project Oxygen and by the Army Research Office under Grant DAAD19-00-1-0466

may be processed separately. Here we consider the simpler task of identifying and modeling a single multi-modal object outlining a methodology which extends our previous work [1] (by incorporating a parametric motion model) and presenting empirical results.

Many useful applications arise when a generative cross-modal appearance model is available. By aggregating the MI measure over explicit sub-regions, a given sound can be classified as coming from one of many possible visual objects.

We exploit a linear manifold model to approximate video appearance. Many models of statistical object appearance have been proposed recently that characterize object appearance. The 'morphable model' of Beymer and Poggio[2] and Jones and Poggio [3], and the related 'active appearance model' of Cootes and Taylor [4] model both the surface variation (texture) and motion (shape) of an object using linear manifolds found by PCA. However, the initialization of these frameworks is challenging in that a set of images must be put in correspondence using manual or semi-automated means. Many models incorporate motion into the appearance model acquisition (c.f. [5]). De la Torre and Black [6] recently demonstrated a robust method for estimating modular PCA models.

This work is of note in that we are interested in cases where information from another domain may make the initialization of these systems trivial. For example, when the object of interest is the only audio source in the scene. In this case we should be able to learn a visual appearance model with no visual initialization, using only audio information *provided* that we have a method of estimating the joint audio-visual signal properties.

Our approach is to enhance a subspace-based audiovisual MI model to include a parametric motion term, such that it can find an optimal warping to simultaneously align observed images over time and optimize MI with the received audio signal. While in theory we could solve this using exhaustive search to try all possible motions, in practice this is computationally infeasible. Augmenting the model with a visual appearance component makes tracking practical.

As we will describe below, we bootstrap a model of object appearance from the initial set of image patches with high pixel variation. Some of the patches correspond to the object of interest and some do not. From these patches we construct an initial subspace matched filter which is used to "track" objects in the scene. During learning we alternate between optimizing the statistics of a smaller audio-visual feature space using the previously estimated object position estimate followed by further refinement of a subspace matched filter.

With this approach we can in principle follow the lips of a speaker when there is momentarily no speech, or when the lips are momentarily occluded but other parts of the face are clearly visible (even if those parts are not individually related to the audio).

In the following sections we describe the approach and related work. We then demonstrate results of the method on the application of finding and modeling a moving noisy object in a dynamic, but otherwise silent, scene. We can define an object by its audio, and automatically build a visual model without any initialization in the visual domain.

## 2. AUDIO-VISUAL STATISTICAL MODEL

We now discuss a generative model of audio-visual data from which we derive approximate inference and estimation algorithms. From the standpoint of pure vision approaches, the assumptions we make are fairly standard with one important distinction, namely that the joint density over dependent audio-video parameters are represented nonparametrically rather than using a simplifying parametric form. The motivation for this choice is that it allows for a richer model of joint audio-visual dependency. The consequence is that we must also derive several novel approximations in order to arrive at a tractable joint tracking and appearance estimation algorithm.

We use superscripts to indicate whether the variable of interest is related to visual appearance (e.g.  $Y^v$ ), audio appearance (e.g.  $Y^a$ ), or audio-visual appearance (e.g.  $Y^{av}$ ). Let  $\{Y_k^v\}$  denote the set of  $N$  observed images (represented as vectors) from a sequence. We model the  $k$ th image in the sequence,  $Y_k^v$ , as a transformation  $T_k$  of linearly combined appearance bases  $\Phi^v$  and  $\Phi^{av}$  plus independent additive Gaussian noise:

$$Y_k^v = T_k \circ \left( [\Phi^v \ \Phi^{av}] \begin{bmatrix} \alpha_k^v \\ \alpha_k^{av} \end{bmatrix} + n_k^v \right). \quad (1)$$

where  $\Phi^v$  and  $\Phi^{av}$  span orthogonal subspaces and  $\alpha_k^v$ ,  $\alpha_k^{av}$  are projection coefficients associated with the respective sets of bases.

Additionally, let  $\{Y_k^a\}$  denote the set of  $N$  associated audio measurements. The  $k$ th audio measurement,  $Y_k^a$ , is a local periodogram (i.e. the magnitude squared of a Fourier transform computed over a window of data) centered in time on the  $k$ th image and windowed. The periodogram samples are also modeled as a linear combination of bases  $\Theta^a$  and  $\Theta^{av}$  with coefficient vectors  $\beta_k^a$  and  $\beta_k^{av}$ , respectively, plus independent additive Gaussian noise:

$$Y_k^a = [\Theta^{av} \ \Theta^a] \begin{bmatrix} \beta_k^{av} \\ \beta_k^a \end{bmatrix} + n_k^a. \quad (2)$$

where  $\Theta^a$  and  $\Theta^{av}$  span orthogonal subspaces and  $\beta_k^a$ ,  $\beta_k^{av}$  are projection coefficients associated with the respective bases. In our model the sets of bases  $\{\Phi^v, \Phi^{av}, \Theta^a, \Theta^{av}\}$  and the transformation  $T_k$  are treated as parameters to be estimated while  $\{\alpha_k^v, \alpha_k^{av}, \beta_k^a, \beta_k^{av}\}$  are treated as random variables.

In Equation (1), the coordinate transformation  $T_k$  accounts for affine motion in the  $k$ th image frame. Coefficients  $\alpha_k^v$  account for iconic changes in the object appearance which are independent of the audio signal, while  $\alpha_k^{av}$  account for iconic changes that are related to the audio. Likewise, the  $\beta_k^a$  account for changes in the audio which are independent of the iconic changes to the image, while the coefficients  $\beta_k^{av}$  account for audio changes which are related. The generative statistical model, depicted in Figure 1, over

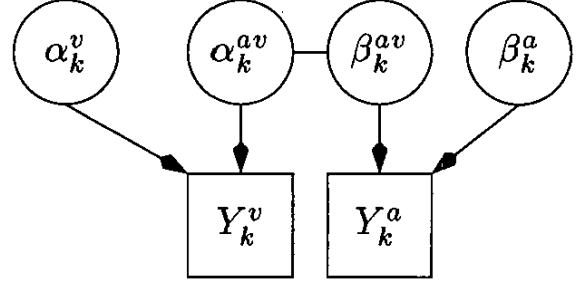


Fig. 1. Generative model of audio-visual appearance (left)

latent variables  $\{\alpha^v, \alpha^{av}, \beta^a, \beta^{av}\}$  and observables  $\{Y^v, Y^a\}$  factors as:

$$p(\alpha_k^v, \alpha_k^{av}, \beta_k^a, \beta_k^{av}, Y_k^v, Y_k^a) = p(\alpha_k^v) p(\beta_k^a) p(\alpha_k^{av}, \beta_k^{av}) \times p(Y_k^v | \alpha_k^v, \alpha_k^{av}, T_k, \Phi^v, \Phi^{av}) \times p(Y_k^a | \beta_k^a, \beta_k^{av}; \Theta^a, \Theta^{av}) \quad (3)$$

where  $p(\alpha^v)$  and  $p(\beta^a)$  are modelled as Gaussian, and  $p(\alpha^{av}, \beta^{av})$  is a Parzen density. It is important to note that joint dependence in the audio and video observations  $Y_k^v$  and  $Y_k^a$  is through the joint density  $p(\alpha^{av}, \beta^{av})$ . In the sequel it is our purpose to simultaneously estimate the parameters  $T_k$ , the appearance bases  $\Phi^v$ ,  $\Phi^{av}$ ,  $\Theta^a$ , and  $\Theta^{av}$ , and a model of the joint density  $p(\alpha^{av}, \beta^{av})$ . Given the indeterminacy of such a task, some care and simplifying assumptions are necessary.

## 3. LEARNING WHILE TRACKING

Having described the generative model, we now discuss a method by which we learn appearance bases  $\{\Phi^v, \Phi^{av}, \Theta^v, \Theta^{av}\}$  and estimate the motion parameters  $T_k$ . We accomplish this via an iterative coordinate optimization procedure (iterations are denoted by the index  $j$ ). A particular challenge in this regard is to learn the joint statistical model  $p(\alpha^{av}, \beta^{av}; \{\alpha_k^{av}, \beta_k^{av}\})$  as well as the associated bases  $\Phi^{av}$  and  $\Theta^{av}$ . Toward that end we decompose the algorithm into three steps, summarized as:

1. **detection:** Given previous (initial) estimates of  $\{T_k, \alpha_k^{av}, \alpha_k^v, \Phi^{av}, \Phi^v\}_{j-1}$ , perform a local search to generate a new set of transformations  $\{T_k\}_j$ .
2. **estimate dependent bases:** Given  $\{T_k\}_j$ , estimate  $\{\Phi^{av}, \Theta^{av}, \alpha_k^{av}, \beta_k^{av}\}_j$ .
3. **estimate independent bases:** Given  $\{\Phi^{av}, \Theta^{av}\}_j$  remove their contribution from the audio and video observations and estimate  $\{\Phi^v, \Theta^a\}_j$ .

### 3.1. Detection

Following the additive Gaussian assumption, conditioned on previous estimates of  $\{\Phi^{av}, \Phi^v, \alpha^v, \alpha^{av}\}_{j-1}$ ,  $T_k$ 's are estimated:

$$T_k = \arg \min_{\{T\}} \left\| T^{-1} \circ Y_k^v - [\Phi^v \ \Phi^{av}] \begin{bmatrix} \alpha_k^v \\ \alpha_k^{av} \end{bmatrix} \right\|_2 \quad (4)$$

using a local search in translation and rotation about  $\{T_k\}_{j-1}$ . Translations are searched efficiently using FFTs while rotation and translation are further refined via downhill simplex [7, 8].

### 3.2. Estimating A/V Dependency

In [1] we presented an approach for learning joint audio-visual statistical models based on a nonparametric estimate of mutual information [9]. We further refined the approach in [10] in which we provided a statistical justification for the method as well as additional regularizing terms related to the learned bases. Since  $\Phi^{av}$  and  $\Theta^{av}$  span subspaces which are orthogonal to  $\Phi^v$  and  $\Theta^a$ , respectively (by construction) we can consider the projection of the observations onto  $\Phi^{av}$  and  $\Theta^{av}$  separately. As such we learn projection matrices  $H_v$  and  $H_a$  which parameterize the linear fusion model:

$$\begin{bmatrix} \alpha_1^{av} & \dots & \alpha_N^{av} \\ \beta_1^{av} & \dots & \beta_N^{av} \end{bmatrix} = \begin{bmatrix} H_v^T & 0 \\ 0 & H_a^T \end{bmatrix} \times \begin{bmatrix} T_1^{-1} \circ Y_1^v & \dots & T_N^{-1} \circ Y_N^v \\ X_1^a & \dots & X_N^a \end{bmatrix} \quad (5)$$

so as to maximize the criterion

$$J = \hat{I}(\alpha^{av}; \beta^{av}) - \lambda_1 H_v^T H_v - \lambda_2 H_a^T H_a - \lambda_3 H_v^T \bar{R}_v^{-1} H_v \quad (6)$$

where 0 indicates a matrix of zeroes with appropriate dimension,  $\hat{I}(a; b)$  is a particular estimate of mutual information [10], and the remaining terms are regularizing priors. The set of bases  $\Phi^{av}$  and  $\Theta^{av}$  are computed as the pseudo-inverse of the matrices  $H_v$  and  $H_a$  respectively.

The inclusion of mutual information in the criterion is motivated by the easily proven inequalities:

$$I(\alpha^{av}; \beta^{av} | \{T_k\}) \leq I(\alpha^{av}; \alpha_*^{av}) \quad (7)$$

$$I(\alpha^{av}; \beta^{av} | \{T_k\}) \leq I(\beta^{av}; \beta_*^{av}) \quad (8)$$

where  $\alpha_*^{av}$ ,  $\beta_*^{av}$  are the "true" latent variables. Consequently, maximizing the mutual information between the estimated latent variables  $\alpha^{av}$  and  $\beta^{av}$  maximizes a lower bound on the mutual information between the estimates and the true underlying latent variables (up to some invertible transformation). Due to the use of mutual information as a similarity criterion, these variables need not have the same form or dimensionality.

### 3.3. Estimating Independent Bases

The final step in each iteration is to estimate the independent bases for audio and video. In order to ensure that these bases span a subspace which is orthogonal to their counterparts in the previous step we begin by removing the contribution of the dependent bases:

$$T_k^{-1} \circ \hat{Y}_k^v = T_k^{-1} \circ \left[ (I - \Phi^{av\dagger}) Y_k^v \right] \quad (9)$$

$$\hat{Y}_k^a = (I - \Theta^{av\dagger}) Y_k^a \quad (10)$$

$\Phi_j^v$  and  $\Theta_j^a$  are taken to be the first principal components of the resulting augmented observations.

### 3.4. Initialization

We use a simple algorithm based on image differences to initialize the algorithm. We begin by apriori choosing an image patch size. Candidate patches are chosen throughout the sequence which have



Fig. 2. Two images from the image sequence (left and center) and the average over all images (right). As can be seen the speaker undergoes significant (primarily translational) motion.

high energy as measured by differences between consecutive images. Some of these patches contain the object of interest while others do not. These patches are used to construct an initial appearance subspace. We'll denote this as  $\Phi_0$ , note that this basis implicitly includes  $\Phi_0^v$  and  $\Phi_0^{av}$ . For experimental purposes, the initial subspace dimension is restricted to 5 (i.e. 5 bases) of size 128x128. These form the basis of an initial subspace filter which is used to locate the object. Thereafter we iterate through the steps outlined above.

## 4. EMPIRICAL RESULTS

In this section we present empirical results using the method described. We use as a test sequence a 15 second audio-video clip of a person speaking and moving their torso from side to side. Most of the motion is translational, however, there is some rotational component when the subject is at the extreme left or right. The subject is speaking during the entire sequence. Additionally there is a computer monitor in the background whose image is also changing. Figure 2 shows two of the images from the A/V sequence when the speaker is approximately at the extreme left and right. The third image in the sequence shows the average over all images. The average image will help to illustrate the effectiveness of the approach in identifying and tracking a moving/speaking person. Additionally, figure 3 shows the difference images between subsequent image frames for the same two images as well as the image of pixel standard deviations. These differences are used to bias the initial choice of appearance bases. It is important to note that changes due to the monitor are in some sense easier to track as the monitor does not change position. At the very least, without biasing the model, one is as likely to learn an "appearance" of the monitor as one is to learn an appearance of the speaker without using the additional information provided by the audio signal.

The dimensions of the image sequence are 360x240 pixels. Images patches, which ultimately comprise the appearance model are 128x128 pixels. As will be seen this is about twice the size needed for encompassing the face of the speaker.

The dimensionality (number of bases) of the combined visual appearance basis was set to five. Two of the dimensions are used to construct the audio-video feature space. They are then combined with the remaining three to construct a subspace matched filter for locating the object in the scene. All learning is done in batch over 450 images from the segment iterating over the steps of the previous section. Anecdotally, we noted that convergence in the beginning is quite slow. Most likely this is due to the lack of image patch alignment to start (which we intentionally avoided). Eventually, the image patches align (via the estimated transforms) at this point convergence is quite quick - approximately 10 iterations. Figure 4 shows the resulting object appearance basis functions. The two at the right correspond to the audio-video feature



**Fig. 3.** Two difference images from subsequent frames in the image sequence (left and center) and image of pixel-wise standard deviations taken over all images(right). Note that black indicates large values while white indicates low values. From difference images we see that changes in the scene due to the monitor are as significant as changes due to the moving speaker. Additionally, as illustrated from pixel-wise standard-deviations, changes due to the monitor are localized, while changes due to the speaker cover a large region.



**Fig. 4.** Images of basis functions learned from the procedure. The two basis images at the right are the result of the cross-modal subspace and exhibit more sensitivity to the lip/chin motion than the remaining three basis images at the left which are more sensitive to the objects static features.

space. They appear to have more sensitivity to lip and chin motion than the three basis images to the left. All five are used to construct a matched subspace filter which is used to locate the object of interest.

Finally, in a repeat of figure 2 we show the same three images *after* after applying the estimated inverse transformation in figure 5. While the single images appear to be centered, it is the average image which is most striking, noting that the facial features come clearly into view.

## 5. DISCUSSION

We have presented a new algorithm for exploiting cross-modal information for purposes of learning an appearance model and estimating the large-scale motion of an *audio-video* object. This method extends our previous work on subspace approaches to learning multi-modal statistical models. While the experimental results are promising, analysis of more complicated scenes will be the subject of future work. Note that we successfully recovered an appearance model of a moving subject using audio-video correspondence in a feature space *without* initializing to the object explicitly. Conceptually, the algorithm is straightforward.



**Fig. 5.** The same two images from the image sequence (left and center) shown in 2 after applying a global transform estimate and the transformed average over all images (right).

## 6. REFERENCES

- [1] John W. Fisher III, Trevor Darrell, William T. Freeman, and Paul. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems 13*, 2000.
- [2] David Beymer and Tomaso Poggio, "Face recognition from one example view," Tech. Rep. AIM-1536, MIT Artificial Intelligence Laboratory, 1995.
- [3] Michael J. Jones and Tomaso Poggio, "Model-based matching by linear combinations of prototypes," Tech. Rep. AIM-1583, MIT Artificial Intelligence Laboratory, 1996.
- [4] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, June 2001.
- [5] Michael J. Black, David J. Fleet, and Yaser Yacoob, "Robustly estimating changes in image appearance," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 8–31, 2000.
- [6] Fernando De la Torre and Michael Black, "Robust parameterized component analysis theory and application to 2d facial modeling," in *7th European Conference on Computer Vision*, Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, Eds., Copenhagen, Denmark, May 28-31 2002, vol. 2353 of *Lecture Notes in computer science*, pp. 653–669, Springer-Verlag.
- [7] J. A. Nelder and R. Mead, "," *Computer Journal*, vol. 7, no. 1, pp. 308–313, 1965.
- [8] H. Press William, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes in C*, Cambridge University Press, 2nd edition, 1992.
- [9] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York, 1991.
- [10] John W. Fisher III and Trevor Darrell, "Probabilistic models and informative subspaces for audiovisual correspondence," in *7th European Conference on Computer Vision*, Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, Eds., Copenhagen, Denmark, May 28-31 2002, vol. 2352 of *Lecture Notes in computer science*, pp. 592–603, Springer-Verlag.