AUDIO-VISUAL FLOW - A VARIATIONAL APPROACH TO MULTI-MODAL FLOW ESTIMATION

Raffay Hamid, Aaron Bobick

GVU Center/College of Computing
Georgia Institute of Technology
{raffay,afb}@cc.gatech.edu

Anthony Yezzi

School of Electrical and Computer Engineering
Georgia Institute of Technology
ayezzi@ece.gatech.edu

ABSTRACT

Just as a motion field is associated to a moving object, an audio field can be associated to an object that can behave as a sound source. The flow field of such a sound source which moves over time would not only have an optical component, but also an audio component; something we call audio-visual flow. In this paper we present a common structure tensor based variational framework for dense audio-visual flow-field estimation. The proposed scheme improves the rank of the local structure tensor by incorporating an audio information channel which is substantially un-correlated from the complementing visual information channel. The scheme allows ascribing weights to individual sensor modalities based on the confidence in their corresponding measurements. Results are presented to demonstrate how combining multiple modalities in our proposed framework can provide a possible solution to temporary full visual occlusions.

1. INTRODUCTION

Information fusion is an important pre-step for many higher-level reasoning tasks. Inference in many interesting problems such as speaker localization in video conferencing, and robot navigation etc, can be made with higher confidence by relying on multi-modal data. We are interested in analyzing the flow fields associated with different sensor modalities as the signal source moves over time. Two of such modalities which are of immediate interest are audio and vision. We refer to the flow fields associated with such scene elements which are mobile and can act as sound sources as $audio{-}visual$ flow fields.

By combining different information modalities, each individual information source may compensate for the weakness of the other. For instance, an object tracker relying only on the visual information may loose the track of the object in case of occlusion. On the other hand a tracker that purely relies on the audio information can perform well as long as the moving object emits sounds. The contribution of this work is to provide a variational framework to measure the vector-valued audio-visual flow fields of scene elements, which can in turn be used as input to higher level perception systems.

Combining audio information channel along with vision also improves the robustness of the framework by improving the rank of the local structure tensor. Adding in more visual channels, e.g. R, G and B in the structure tensor framework may improve the tensor rank; however, these channels are extremely well-correlated due to the photometry of the scene. Therefore in most of cases, the improvement in tensor rank is mainly due to the added noise effect. Hence there is a need to add in some information that is highly un-correlated from the visual information channels. This purpose is served by adding in the audio information.

We begin by briefly describing the notion of optical flow from a total least-square solution perspective as described in [1]. We then show how the technique proposed in [1] can be extended to vector valued image sequences and how doing this will improve the rank of the local structure tensor. We then describe how the probability of the presence of an audio source is calculated and mapped to the image coordinate system. Once both the audio and the video signals have been converted in a coherent framework, we explain how these two information sources can be fused in a generalized $audio{-}visual$ structure tensor, the eigen analysis of which can be used to estimate the audio-visual scene flow. Finally, we present results to demonstrate how our framework provides a possible solution for temporary full occlusion. Fig 1 shows the overview of the underlying motivation of our proposed framework.

1.1. Related Work

Optical flow estimation is one of the classic problems in computer vision and has been deeply studied over more than two decades(see e.g. [2] [3] [4]). More recently, Haussecker et al [5] proposed a total least square solution to this problem which is equivalent to a tensor representation of the spatio-temporal image structure. Such a representation varies both the spatial as well as the temporal flow vectors and hence leads to a more precise solution. We base our frame-

0-7803-8554-3/04/$20.00 ©2004 IEEE.

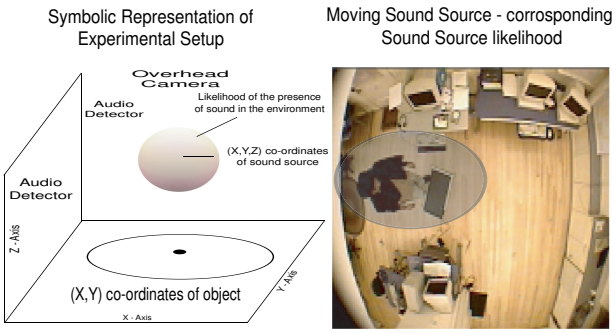

Fig. 1. (a) Symbolic diagram showing the setup of the audio-visual data capture experiment. The visual data is captured through an over-head camera while an array of microphones are used to estimate the 3-D world co-ordinates of the sound source. (b) A sample frame taken from the over head camera. The blob over-laid upon the person is the 2-D image projection of the 3-D probability density representing the presence of a sound source in the ambient environment.

work on the work presented in [5] and extend it to vector valued image sequences for multi-modal data.

In the past there have been numerous attempts to combine audio and video data to improve higher level inference of scene activities (see e.g. [7]). While these techniques produce useful results, most of them rely upon combining different modalities at a higher level, and hence are somewhat domain specific. On the other hand, our proposed framework attempts to fuse these modalities at image level, and hence can be used to solve a broad spectrum of higher level perception problems.

2. TENSOR-BASED VISUAL FLOW-FIELD ESTIMATION

The displacement of gray value structures within consecutive images of a sequence yields inclined image structures with respect to temporal axis of spatio-temporal images. The relation between the orientation angle and the optical flow is given by

$$f = [tan\varphi_x, tan\varphi_y] \tag{1}$$

where $f = [f_x, f_y]$ denotes the optical flow on the image plane and the angles φ_x and φ_y define the angles between the plane normal to the lines of constant grey value along the x and y axes respectively. This is illustrated in figure 2. This basic property of spatio-temporal images allows us to estimate the optical flow from a 3D orientation analysis, searching for the direction of constant gray value in the $\mathbf{x}-t$ space. Let $g(\mathbf{x})$ represent a spatio-temporal image sequence where $\mathbf{x} = (x_1, x_2, x_3)$ where x_3 represents time. The local 3D-structure tensor as defined in [1]can be written as

$$J(\mathbf{x}) = \int h(x - x^{'}) \nabla g(\mathbf{x}^{'}) \nabla g^T(\mathbf{x}^{'}) d^W x^{'} \tag{2}$$

The components of J can be given as

$$J_{pq} = \int h(x - x^{'}) \frac{\partial g(x^{'})}{\partial g(x_p)} \frac{\partial g(x^{'})}{\partial g(x_q)} d^W x^{'} \tag{3}$$

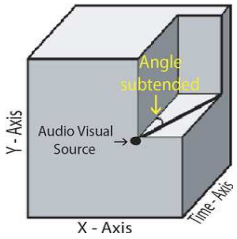

Fig. 2. Visualization of Spatio-Temporal brightness pattern created by a moving object. The figure shows the angle subtended by the motion trajectory along the x-axis φ_x. In a similar way the motion trajectory subtends angle φ_y along the y-axis

The information within a local neighborhood around the central point x is weighted by a window-function $h(x-x^{'})$. The matrix formulation of Eq.2 can be written as:

$$J(\mathbf{x}) = \begin{pmatrix} g_x^* g_x^* & g_x^* g_y^* & g_x^* g_t^* \\ g_x^* g_y^* & g_y^* g_y^* & g_y^* g_t^* \\ g_x^* g_t^* & g_y^* g_t^* & g_t^* g_t^* \end{pmatrix} \tag{4}$$

where g^* represents the smoothed version of the individual components of J. The eigen vectors of J give the local orientations, and the corresponding eigen values denote the local gray-level variations along these directions.

2.1. Eigen Analysis

An eigen analysis of the structure tensor corresponds to a total least squares fit of a locally constant displacement vector field to the intensity data. Let μ_i denote the eigen values of J where $i \epsilon \{1, 2, 3\}$, and μ_i are sorted in a descending order. Let e_i represent the corresponding eigen vectors of J. By analyzing the rank of the matrix, four different cases of spatio-temporal structures can be distinguished:

i- Rank(J) = (0,3): No apparent linear motion is observed.
ii- Rank(J) = 1: An already oriented image structure moves with a constant velocity. This is the well known $aperture\ problem$ in optical flow computation. Only one of the three eigen vectors has an eigen value larger than zero. This eigen vector $e_l = (e_{l,x}, e_{l,y}, e_{l,t})$ points normal to the plane of constant grey value in 3D space and can be used to compute the normal optical flow as:

$$f = \frac{e_{l,t}}{(e_{l,x}^2 + e_{l,y}^2)^{1/2}} \tag{5}$$

iii- Rank(J) = 2: An isotropic grey value structure moves with a constant velocity. No aperture problem is present in the spatio-temporal neighborhood. The orientation of the 3D iso-grey-value line yields the two components f_1 and f_2 of the optical flow. The flow vector in this case can be computed as:

$$f = (\frac{e_{s,x}}{e_{s,t}}, \frac{e_{s,y}}{e_{s,t}}) \tag{6}$$

3. AUDIO SOURCE LOCALIZATION

For the sake of technical completeness, we would briefly go over the notion of audio source localization. Let us assume

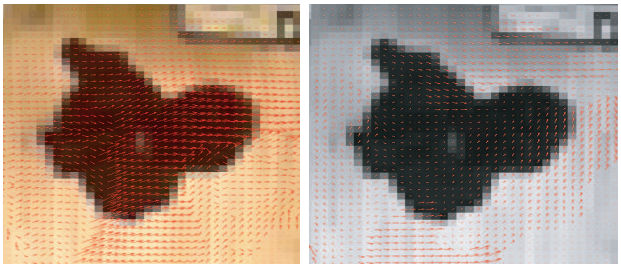

Fig. 3. (a)Vector Valued image overlayed by the optical flow-field. As can bee seen, the optical flow field suffers less from aperture problem. (b) Gray-Scale equivalent of the vector valued image. As can be noticed, the flow-field suffers more from aperture problem.

we have an array of microphones in our 3D environment. Given a single source of sound that produces a time varying signal $x(t)$ each microphone in the array will receive the signal $m_i(t) = \alpha_i x(t-t_i)+n_i(t)$ where i is the microphone number, t_i is the time it takes sound to propagate from the source to microphone i, and $n_i(t)$ is noise signal present at microphone i. The Time Delay of Arrival (TDOA) is defined for a given microphone pair (i, k) as $D_{ik} = t_i - t_k$. The idea is to determine D_{ij} for some subset of microphone pairs, and then finding the least mean square solution for the sound source. The Fourier Transform of the captured signal can be expressed as:

$$m_i \longleftrightarrow M_i(w) = \alpha_i X(w) e^{-jwt_i} + N_i(w) \tag{7}$$

where $X(w)$ is the Fourier Transform of the source signal $x(t)$.

The cross correlation of $m_i(t)$ and $m_k(t)$ can be given as:

$$R_{ik}(\tau) = \int m_i(t) m_k(t-\tau) dt \tag{8}$$

The frequency domain representation of $R_{ik}(\tau)$ can be given as:

$$R_{ik}(\tau) \longleftrightarrow S_{ik}(w) = M_i(w) M^*(k) \tag{9}$$

$S_{ik}(w)$ can be approximated using eq 7 as:

$$S_{ik}(w) = \alpha_i \alpha_k |X_i(w)|^2 e^{-jwD_{ik}} \tag{10}$$

Thus D_{ik} can be found by evaluating:

$$D_{ik} = max(R_{ik}(\tau)) = max(\mathcal{F}^{-1} S_{ik}(w)) \tag{11}$$

where $\mathcal{F}^{-1}$ represents the inverse Fourier Transform. The location for the sound source is a point $\mathbf{p}$ that satisfies the set of associated parametric equations:

$$\left\{ \begin{array}{c} \frac{d(p-m_1)-d(p-m_2)}{V_{sound}} = D_{12} \\ \cdot \\ \cdot \\ \frac{d(p-m_i)-d(p-m_k)}{V_{sound}} = D_{ik} \end{array} \right\} \tag{12}$$

where $d(p-m_k)$ represents the euclidian distance between the sound source and the microphone. Thus a set of three $D'_{ik}s$ uniquely specify the coordinates of the source. For sets of four or more D_{ik} a solution may only exist in a least mean square sense.

4. AUDIO-VISUAL FLOW-FIELD ESTIMATION

In this section we first explain how we can incorporate R, G and B channels to compute optical flow for vector valued images. We then move on to explain how the audio and video information can be combined to compute audio-video flow field.

4.1. Optical Flow Estimation for Vector-Valued Images

Thus far, the focus of calculation of optical flow for gray scale images $I : \Omega \subset \mathbb{R}^2 \longrightarrow \mathbb{R}$ has been mainly due to the reasons of computational efficiency. However, thanks to the improvement of processing speed, it seems logical to extend the computation of optical flow on vector-valued images $I : \Omega \subset \mathbb{R}^2 \longrightarrow \mathbb{R}^n$. Due to the availability of more data points corresponding to a particular point in the image space, as compared to the grey scale case, vector valued optical flow can better take care of the $aperture\ problem$. In this section we first formulate an extension of method described in section 2.1 to vector valued image sequences. For vector-valued images, $\nabla g(\mathbf{x})$ as defined in eq 2 can be extended as:

$$\nabla g(\mathbf{x}) = \begin{pmatrix} R_x & G_x & B_x \\ R_y & G_y & B_y \\ R_t & G_t & B_t \end{pmatrix} \tag{13}$$

The structure tensor $J(\mathbf{x}) = \nabla g(\mathbf{x}) \nabla g(\mathbf{x})^T$ is still of order 3×3, however, it now contains the variational information in vector-valued spatio-temporal hyperspace about the immediate neighbors of each pixel. It is important to note here that the formulation of the structure tensor as given in eq 4 is inherently of rank 1. Only the fact that each pixel is weighted by its neighbor in a non-linear fashion, i.e. by using an exponentially decreasing spatial weight function, increases the tensor rank $J(\mathbf{x})$. In case of vector valued image sequences however, since we are using more information for every pixel point in the image space, therefore there is a higher likelihood that the structure tensor will be of fuller rank even before the affect of the neighboring pixels is brought into picture. Smoothing improves the rank only further. Since the determination of the type of motion, and hence the flow computation are based on the rank of $J(\mathbf{x})$, this improvement of the rank significantly improves the optical flow results. Figure 3 shows the comparisons when the algorithm described in Sec. [1] was applied on the gray scale version of the image and the result of our proposed extension to the vector valued image sequence. As can be seen, the optical flow field in case of vector-valued image suffers much less from the $aperture\ problem$ as compared to the gray scale case.

4.2. Audio-Visual Structure Tensor

The framework presented in Section 3 can be extended to a probabilistic domain, i.e. assuming a gaussian noise in our measurement, the estimation of the sound source location can be considered as a 3D gaussian distribution in the

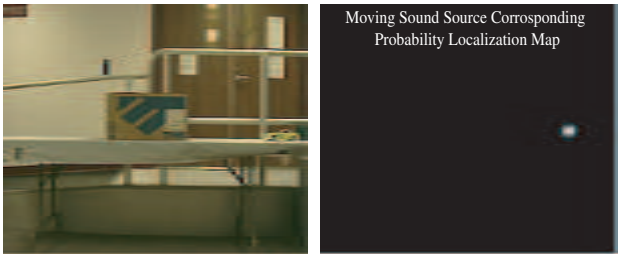

Fig. 4. (a) Sample frame of a moving car acting as a sound source. (b) Corresponding probability source localization map.

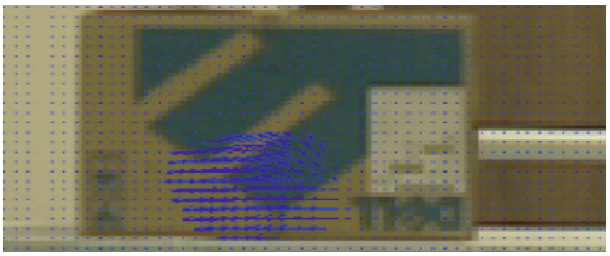

Fig. 5. Sample frame of a fully occluded moving car. The proposed framework leverages the audio channel to estimate the flow field of the fully occluded moving sound source.

3D world co-ordinate system. For our purposes, we are currently simulating the audio probability distribution around the moving body. In the future we intend to implement audio source localization algorithm as described in Sec. 3 for cross-validation purposes.

This distribution is projected on the image plane using the projective transformation equation obtained from camera calibration. Thus we have now found the 2D projection of the 3D density for the estimation of the audio source location. A sample frame and the corresponding probability map is shown in fig 4. We now extend the modified expression for the vector-valued spatio-temporal gradient of the image sequence as defined in eq 13 such that it can now incorporate the audio field as well. This can be given as:

$$\nabla g(\mathbf{x}) = \begin{pmatrix} R_x & G_x & B_x & S_x \\ R_y & G_y & B_y & S_y \\ R_t & G_t & B_t & S_t \end{pmatrix} \tag{14}$$

where S_x, S_y and S_t represent the spatio-temporal first-order derivative of the sound localization probability distribution. Since we know the analytical form of this probability distribution, we can compute the fourth column of the matrix given in eq14 in an analytical fashion. However for now we resort to numerical methods to compute it by finding the first-order difference of the sound localization probability image. Since the intensity of the sound localization image is in direct proportionality to our belief in the presence of the sound source at that location, we can also modify eq 14 to obtain a more robust estimate of the audio-visual flow field. Thus eq 14 can be modified as:

$$\nabla g(\mathbf{x}) = \begin{pmatrix} R_x & G_x & B_x & \alpha S_x \\ R_y & G_y & B_y & \alpha S_y \\ R_t & G_t & B_t & \alpha S_t \end{pmatrix} \tag{15}$$

In Eq 15 α represents the value of the 2-D image-plane projection of the 3-D spatial probability distribution of the audio source location. We apply our proposed audio-visual flow-field estimation framework, to solve the problem of full visual occlusion. This is illustrated in figure 5. As can be seen, even when the moving sound source is behind the obstacle and is fully occluded, the audio information channel still proves meaningful information and the resulting flow field gives a decent estimation of the position of the moving sound source.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have described a common variational framework for estimating the audio-visual flow field associated to a moving audio source. We have first described a procedure of computing the optical flow of vector valued images which improves the rank of the local structure tensor. We have then presented a way of incorporating audio flow field in a unified variational framework. Results are presented that indicate that the proposed framework can be useful for multi-modal signal fusion which can in turn be applied to solve various perceptual problems. These may include understanding of the layered representation of the scene and tracking under full occlusion. In the future we would like to apply our low level audio visual features to track an object of interest. Finally, the current work is under the assumption of the presence of only one mobile sound source. In the future we would like to investigate how this work can be extended to multiple mobile sound sources.

6. REFERENCES

[1] H.Spies, and H. Scharr, "Accurate Optical Flow in Noisy Image Sequences". ICCV, 2001.

[2] B. K. P. Horna and B. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185 - 204, 1981.

[3] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", Proceedings of the 7th Intel. Conf. on Artificial Intelligence, 1981, pp. 674–679

[4] M. J. Black and P. Anandan, "The robust estimation of mutiple motions: Parametric and piecewise-smooth flow fields." CVIU, 63(1): 75 - 104, 1996.

[5] Haussecker, Spies and Janhe, "Tensor-based Image Sequence Processing Techniques for the Study of Dynamical Processes", International Symposium on Realtime Imaging and Dynamic Analysis, Hakodate, Japan, June 2-5, 1998.

[6] C. Wang and M. Brandstein, "Multi-source face tracking with audio and visual data", In IEEE Intl. workshop on Multimedia Signal Processing, 1999.

[7] M.J. Beal, H. Attias and N. Jojic, "Audio-Video Sensor Fusion with Probablistic Graphical Models". ECCV 2002.