Adaptive Space-Time Structural Coherence for Selective Imaging

David Gibson
Computer Science
University of Bristol
Bristol, UK
David.Gibson@bristol.ac.uk

Neill Campbell
Computer Science
University of Bristol
Bristol, UK
Neill.Campbell@bristol.ac.uk

***Abstract*—In this paper we present a novel close-to-sensor computational camera design. The hardware can be configured for a wide range of autonomous applications such as industrial inspection, binocular/stereo robotic vision, UAV navigation/control and biological vision analogues. Close coupling of the image sensor with computation, motor control and motion sensors enables low latency responses to changes in the visual field. An image processing pipeline that detects and processes regions containing space-time structural coherence, in order to reduce the transmission of redundant pixel data and stabilise selective imaging, is introduced. The pipeline is designed to exploit close-to-sensor processing of regions-of-interest (ROI) adaptively captured at high temporal rates (up to 1000 ROI/s) and at multiple spatial and temporal resolutions. Space-time structurally coherent macro blocks are detected using a novel temporal block matching approach; the high temporal sampling rate allows a monotonicity constraint to be enforced to efficiently assess confidence of matches. The robustness of the sparse motion estimation approach is demonstrated in comparison to a state-of-the-art optical flow algorithm and optimal Baysian grid-based filtering. A description of how the system can generate unsupervised training data for higher level multiple instance or deep learning systems is discussed.**

***Index Terms*—Close-to-sensor processing, low latency processing, feature analysis, sub-pixel tracking.**

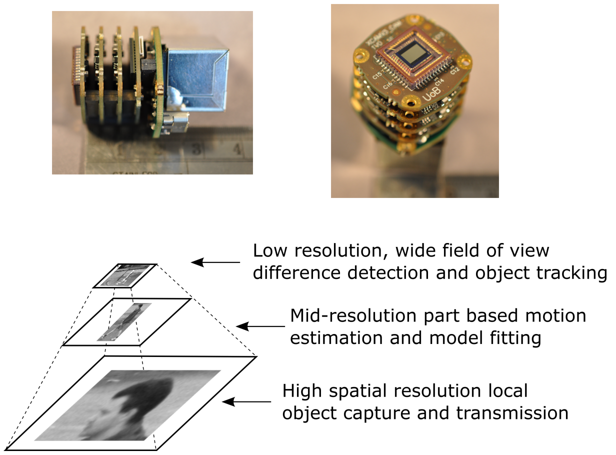
Low resolution, wide field of view
difference detection and object tracking
Mid-resolution part based motion
estimation and model fitting
High spatial resolution local
object capture and transmission

Fig. 1. The close-to-sensor design presented in this paper (top). The configuration includes a highly programmable image sensor, multi-core concurrent processors, 9 degrees of freedom motion sensing and weighs under 25 grams. An illustration of the multi-scale adaptive region selection and person tracking pipeline (bottom). Tracking can be performed at lower resolutions; only high resolution head pixels need to be transmitted for identification.

I. Introduction

Object detection and tracking are critical mechanisms for understanding the changes that occur in a dynamic scene. However, the act of image capture and the consequent mapping of the continuous visual world into a spatially and temporally quantised digital representation introduces artifacts that affect motion estimation and scene interpretation. A standard camera captures images at fixed frame rates between 25 and 60Hz; as an object moves across the scene significant spatial displacements can occur causing large frame-to-frame differences. For complex motions, large spatial displacements can introduce non-linear space-time transformations increasing the chance of errors in motion estimations.

Several further confounding observations can be made regarding the standard image capture and processing pipeline. Firstly, all pixels are captured regardless of the information content or the task being undertaken; large amounts of redundant pixel data are captured, encoded and transmitted, only to be discarded at the early stages of image processing. Secondly, large amounts of data are generated which require expensive memory and transmission pipelines along with considerable computer processing power. Thirdly, as data amounts increase, image transmission and pixel processing times introduces latency. This is a particular problem for realtime systems such as robotics and UAV navigation where rapid reaction rates are critical. One obvious way around this problem is to increase the frame rate. However, within a standard capture architecture this just acts to increases all of the above difficulties; ultimately, standard imaging does not scale well in the spatio-temporal domains.

In the natural world, extremely efficient biological vision systems have evolved, stabilising visual perception for avoiding collisions, communicating, navigating, etc. However, these well adapted systems can become confused and disorientated under certain conditions suggesting that expectations about the visual world and the speed at which events occur are to some extent learnt [1]. Additionally, specific adaptations such as the peripheral and fovea in humans enables a visual system capable of monitoring and attending to the visual world at

978-1-5386-3534-6/17/$31.00 2017 IEEE

multiple spatio-temporal resolutions while preparing to attend to the next most important event.

In this paper we present a novel architecture that has the potential to overcome many of the issues highlighted above. The architecture is biologically inspired but uses off-the-shelf-components in a highly parallel, compact design. The system is adaptable and low cost by virtue of modern image sensor and processor technologies and the fact that redundant pixel data is not captured, transmitted, stored or processed. As such, a visual scene is observed at multiple spatial and temporal scales in order to filter only objects or object parts that are of interest, Figure 1.

II. Background

Original investigations into autonomous low latency vision and motor control were carried out in [2] where it was found that latency between vision processing, motor control and inertial lag limited the effectiveness of a binocular robotic system. In [3] low latency *sensory attention*, vision processing was investigated using a bespoke VLSI architecture. A major difficulty of such systems is that off-the-shelf components do not have the required levels of performance or that build-your-own silicon hardware becomes prohibitively expensive.

One area of contemporary research is the use of neuromorphic engineering techniques to implement circuits that respond to optical flow or other changes in a scene. In [4] a very low latency, event driven device is presented. While changes are rapidly detected and processed via binary spikes, no pixel intensity information is captured. In [5] pixels are processed on the focal plane to remove redundant data and output only salient areas of pixels; by doing so energy consumption is quantifiably reduced. Recently the Chronocam has been introduced. There are few technical details available; it seems to combine properties of the two aforementioned devices and ideas presented in this paper. Each of this technologies offers interesting application potential but none offer the low cost, richness of functionality and programmability of the proposed device.

In [6] and [7] it was shown that adaptive ROI sampling using a novel combination of off-the-shelf components is possible. A natural extension to this adaptive temporal processing pipeline is to sample pixels at variable spatial resolutions. Disregarding the physical repositioning of the lens, such a capability is made available by the programmable digital zoom, or sub-sampling, mechanism of modern image sensors. A major advantage of having advanced sensor surface sub-sampling is that wide fields of view can be attended to, albeit at a low spatial resolution. Salient objects at low resolution can be re-sampled at higher spatial resolutions and high temporal resolutions for further, more complex, processing while keeping the data transfer rates low. Adding a form of foveation to an adaptive camera system enables the investigation of more complex computational responses to moving visual stimuli. In particular, the question of how many, where and how often pixels should be sampled to successfully complete a particular task is not well understood.

Optical flow and motion estimation are fundamental components for dynamic scene interpretation and have been extensively investigated. In this work we are particularly interested in frame-to-frame techniques that do not require the storage of multiple frames or significant temporal processing. Traditional differential filter based approaches have been proposed for smooth optical flow and tracking of interesting points [8]–[10], while the motion estimation for compression relies on the efficiencies of block matching. The accuracy of these approaches can be dramatically affected by a number of factors, including noise, discontinuous variation in appearance or the absence of smooth pixel gradients. Pixel changes in dynamics scenes can be extremely complex and increasingly sophisticated algorithms have been developed to attempt to improve accuracy [11], [12]. However, such approaches require significant amounts of computation and are not generally well suited to realtime processing. In [13] an extension to block matching is introduced whereby only areas that exhibit differences are analysed for local matches within a light-weight probabilistic sub-pixel resolution framework.

In this paper we present an adapted temporal block matching approach that can guarantee, with a measure of certainty, that a point of interest on an object is a reliable measure of motion *and* appearance. By capturing a wide field scene at low spatial resolution, redundant areas of pixels can be filtered out. Areas that exhibit change can be captured at higher spatio-temporal resolutions and subjected to the proposed temporal block matching, i.e. does the local area's appearance *and* motion appear to change in a coherent manner. As the temporal sampling rate is expected to be higher than an object's rate of motion local intensity variations become smaller and more linear and in turn improve the robustness of motion estimation.

The rest of this paper consists of a description of the hardware architecture followed by a detailed description of the proposed motion estimation technique. A discussion of results and the potential applications is followed by the conclusions.

III. Hardware

The modular design achieves its low complexity and cost due to the fundamental idea that most pixels are redundant and contain little or no useful information. The close-to-sensor design combined with state-of-the-art image sensors enables a wide field of view to be analysed for changes using on-sensor image sub-sampling. Areas with changes are re-sampled at higher spatial resolutions; pixels are processed at read-out and discarded if not of interest. In this manner, whole, high resolution images never exist; low resolution images and higher resolution regions of interest are sampled and processed immediately. The result of image processing at various scales are used to update the image sensor configuration according to the nature of the visual stimuli and the task being carried out. As there are no image buffers, high frequency and costly memory storage and transmission is avoided.

The processors used are dual core XMOS XS1 architecture, each running at 100MHz with four concurrent threads. Each core has 64KB of RAM, and as such the design is very much a real-time system as there is no memory to store, or time (in clock cycles) to process large amounts of pixel data. The processor boards can be stacked to create a pipeline consisting of as many processors as needed, up to eight have been tested. An additional feature is that there is no operating system and timing is deterministic and event driven. The processors are programmed using a C-like language and enable concurrent image capture and processing, for example, multiple threads might handle each of the following concurrently; pixel read-in, pixel processing, sensor reprogramming, data transmission, motor driving, etc. Figure 2 show a schematic for a typical configuration. A typical camera stack configuration consists of an image sensor board, two processor boards, a debug board and an ethernet socket add-on, as in Figure 1

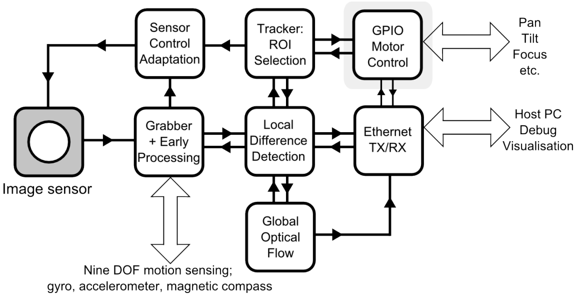
Sensor Control Adaptation
Tracker: ROI Selection
GPIO Motor Control
Pan Tilt Focus etc.
Image sensor
Grabber + Early Processing
Local Difference Detection
Ethernet TX/RX
Host PC Debug Visualisation
Global Optical Flow
Nine DOF motion sensing;
gyro, accelerometer, magnetic compass

Fig. 2. Processing is distributed across threads/cores; processing closer to the sensor delivers lower latency. Multiple sptio-temporal resolutions are supported, for instance low resolution event detection and region-of-interest (ROI) processing. In the binocular configuration a bank of PID controllers drive pan/tilt motors based on visual feedback.

The image sensors are the E2V monochrome Ruby and colour Sapphire, the latter with a resolution of 1600x1200 pixels. The Ruby consists of; 1.3M pixels, 1280x1024, 5.3 square pixels with micro-lensing, an optical format of $1/1.8$, global and rolling shutters and 200mW power usage. A lightweight driver provides full access and control of the sensor registers, such as, windowing (region-of-interest positioning), sub-sampling, binning (averaging), frame rate (exposure time), etc. Region-of-interest read out rates of up to 1500fps are possible for a 64x40 ROI while maintaining a high level of light sensitively and low noise levels. The image sensor headboard includes GPIO for an external trigger and strobe plus a 9 DOF imaging sensing capability; a gyro, accelerometer and e-compass.

IV. Temporal Block Matching

For the presented hardware to process dynamic scenes or ego motion, stable robust motion estimation is required. As whole images do not exist and compute resources are limited, motion estimation has to be applied to sub-sampled ROI in an efficient online manner. Block matching is used extensively in video coding for estimating the motion of macro blocks from one frame to the next. Given a block of pixels in the first frame, the most similar block is searched for within a local area in the next frame. Here we demonstrate the benefits of image sampling at rates that are faster than the object motion; rather than searching over many locations in the spatial domain, we sample a single location though time. A popular metric for computing matches is the mean absolute difference, as in:

$$MAD = \frac{1}{N^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} |C_{ij} - R_{ij}| \qquad (1)$$

where R is the reference macro block centred at x^r, y^r in image one at time $t = 1$ and C is the current macro block centred at x^c, y^c in image two at time $t = 2$. The search area for a good match is decided by the search parameter, p, where p is usually the number of pixels on all four sides of the corresponding macro block in the previous frame. We introduce an extension to Equation 1 that exploits the high temporal sampling rates achievable with the close-to-sensor host architecture. For general, natural scenes, high temporal sampling increases the probability that any frame-to-frame motion is less than one pixel. This allows an assumption of *linearity of motion*, constraining the search parameter to $p = 1$ and for block matching to proceed through time:

$$min(MAD_{t=1}^{T}) \quad s.t.\ (x_t^c, y_t^c) \in [0,1], x = y \neq 0 \qquad (2)$$

Where T is some limit on the number of temporal samples to search over. Equation 2 searches through time over a spatial extent of $p = 1$ until a minimum is found, this search is effectively a sub-pixel matching over the point spread function (psf) of the photon capture process and optics. As such a further *fundamental* constraint can be introduced; that of a monotonic decrease in the MAD block match in the true direction of motion. Within a degree of error there are three conditions that defy this constraint; no-motion-no-texture, edge motion faster than the sampling rate or significant illumination fluctuations occurring close to the sampling rate. If the pixel based block match does not monotonically decrease in any of the single pixel offsets of the $C_t, t > 1$ the block cannot be reliably tracked.

A. Temporal Block Match Tracking

As matching pixels can be prone to noise and given the observation that many significant frame-to-frame differences occur at object boundaries such as edges, it would make sense to investigate other, perhaps more appropriate, feature spaces. Other than noise, frame-to-frame differences are caused by the movement of high contrast edges; features that encode these properties, as well as, scale and orientation include; Haar-like features and log-Gabors. One reason to match in higher dimensional feature spaces is that an implicit encoding of local appearance could provide robustness as well as forming the basis for higher level learning.

Macro block matching though time can proceed until a minimum is found at a one pixel offset in one particular direction; this is straight forward with highly correlated one dimensional pixel or derivative edge values. However, when higher dimensional de-correlated feature spaces are matched, block differences becomes more complicated and a simple difference metric may not represent a good measure of similarity. In order to accommodate higher dimensional features spaces, block matching is tracked using optimal grid-based filters in the difference space. As the sampling rate is higher than the pixel based rate of motion of an object, discrete states, $x_t^i, i = 1, ..., N$ can be defined as the eight, N, directions of one pixel grid offsets. Equations 3-5 represent the posterior, prediction and update of the optimal grid-based tracker, as described in [14].

$$p(x_{t-1}|z_{1:t-1}) = \sum_{i=1}^{N_s} w_{t-1|t-1}^i \delta(x_{t-1} - x_{t-1}^i) \tag{3}$$

$$p(x_t|z_{1:t-1}) = \sum_{i=1}^{N_s} w_{t|t-1}^i \delta(x_t - x_t^i) \tag{4}$$

$$p(x_t|z_{1:t}) = \sum_{i=1}^{N_s} w_{t|t}^i \delta(x_t - x_t^i) \tag{5}$$

Where z_t is the sum of all the feature dimensions that reduce the difference in each direction or state. The feature dimensions are the filter responses and correspond to a weighted approximation of appearance; feature dimensions that have a strong response to a local pattern and move so as to reduce block differences lead to higher weights in the tracker. The delta functions serve as a measure of confidence in any given state, i.e. how many dimensions support a particular hypothesis or direction of motion.

Tracking of the states proceeds until $p(x_t|z_{1:t})$ falls below a threshold or less than half of the dimensions support a particular hypothesis or direction of motion; the last state to be tracked with the largest probability becomes the final motion estimate. Initialisation of the grid based filter states is set to $x_{t=0}^i = d$, where d is the dimensionality of the feature space, i.e. all states are equally likely and all dimensions are reducing block difference.

V. Experiments

To develop and test the temporal block matching approaches, a series of 1000fps monochrome sequences of people walking were captured using a Mikrotron EoSens CL 1362 high speed camera at VGA resolutions. The high resolution images allow an off-line simulation of the sub-sampling capability of the close-to-sensor computational camera system. Even with one to two second sequences lighting conditions can be seen to change rapidly and there are vehicles moving in the background of some clips.

In the first instance, frame-to-frame differencing is carried out at the highest level of sub-sampling (largest pixel size) with a threshold fixed at 8, a value empirically chosen to be just above pixel noise levels. A sliding window is applied to the integral image of the differences; the sub-window with the largest magnitude of differences is used to calculate the coordinates of an ROI at a lower level of sub-sampling (smaller pixel size). A series of up to T ROI are grabbed; as each ROI is grabbed consecutive threads along the pipeline apply filters, extract edges/features and step through the temporal block matching/tracking. When a motion has been estimated the camera front end is put back into the highest level of sub-sampling and the above process repeated.

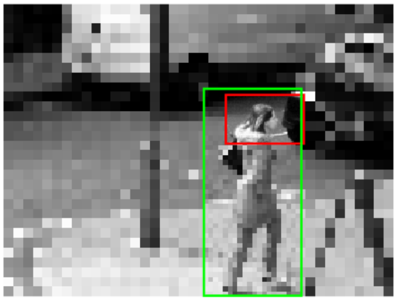

Fig. 3. A person being tracked and sampled at multiple resolutions. The green rectangle represents a prior model of shape and scale, differences under which should be created by something with human proportions. The red rectangle is the expected position of the head within the shape model.

In Figure 3 the test system is illustrated; the green rectangle is the person shape and scale prior; differences under this prior are tracked at the coarsest resolution, motion estimates are computed at the finer resolution samples under the prior and the finest pixels within the head prior (red rectangle) are transmitted.

The multi-scale capture approach can be combined with multi-dimensional feature spaces to improve robustness of tracking and act as a weak model for approximating local appearance. Two feature spaces were investigated; a ten dimensional Haar-like feature set at scales of 4x4 and 8x8 pixels and a log-Gabor filter set using the default settings of [15] to give a 24 dimensional feature set (4 scales, 6 orientations). Both types of features were applied at the whole image scale and separately to the ROI in order to assess the influence of edge effects introduced by ROI sampling.

In this paper, the bounding box person prior is tracked using a sub-pixel-motion optimal grid based tracker; the finer detailed motion estimates of parts is used to illustrate the potential of the proposed hardware configuration with respect to improving robustness.

VI. Results

In Figure 4 the temporal block process is illustrated with plots of similarity in each direction of the single pixel offset search window. The search is plotted for a pixel error and seven of the Haar-like filter responses. For the one dimensional pixel search Equations (1) and (2) are used; for the n-dimensional features spaces the equations of section IV-A are used. In Figure 4, bottom right, the difference is cause by a strong edge moving from left to right; it can be clearly seen for the pixel space and most of the feature dimensions that a one pixel motion creates a minimum in appearance difference at around four temporal frames. Searching in the wrong direction causes an increase in error and some of the features, four and five in particular do not respond well to the underlying appearance.

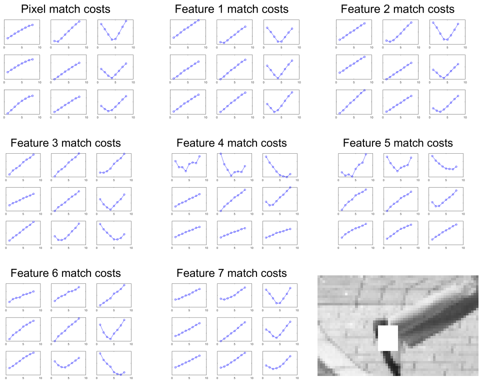

Fig. 4. Temporal block matching for an ROI in pixel and Haar-like feature dimensions. The 3x3 plots show the directions of the single pixel offset search window. No motion is included to illustrate the detection of frame-to-frame differences.

As the person moves differences are detected and ROI from various parts of the person are selected for motion estimation, occasionally the slowly moving right-to-left van in the background is detected. Two estimation approaches were tested; a monotonically decreasing error constraint and an unconstrained grid based filter. For comparison feature extraction was carried out using whole images or the current ROI. In Figure 5 a plot of error for the different motion estimation approaches is shown. The errors are given as the average differences from a reference optical flow estimation of [11]. The spatial x direction of motion estimation has a low error and most of the error comes from incorrect temporal estimation for single pixel motion. To some extent this is reflected in Figure 4 where ambiguity in minima of the temporal domain for single pixel motion can be seen.

It can be seen in Figure 5 the most accurate approach when compared to the reference base-line of [11] is the grid based filter approach in a whole image 24 dimension log-Gabor space. However, it can also be seen that the grid based filter approaches in 10 dimensional Haar-like space are comparable. Using ROI based filter responses, as opposed to full size image responses, indicates that boundary effects increase the error in motion estimation. Next best are the monotonically constrained descents in Haar-like and edge space. Interestingly, the edge space search is almost as good as the 10 dimension Haar-like space while being much more efficient to compute. Notably the pixel based match is the poorest of all indicating noise sensitivity and incoherent pixel based appearance.

In Figure 6 it can be seen that errors in motion estimation increase the slower the motion is. Intuitively, it could be considered that under natural variable lighting conditions the larger the temporal gap between two views of an object the more chance there has been for its appearance to have changed. Additionally the motions in Figure 6 cluster into two groups; those of the fast moving foreground person and the slower moving vehicle in the background. Figure 7 shows a snapshots of ROI tracking over longer periods of time for one of at set of image sequences.

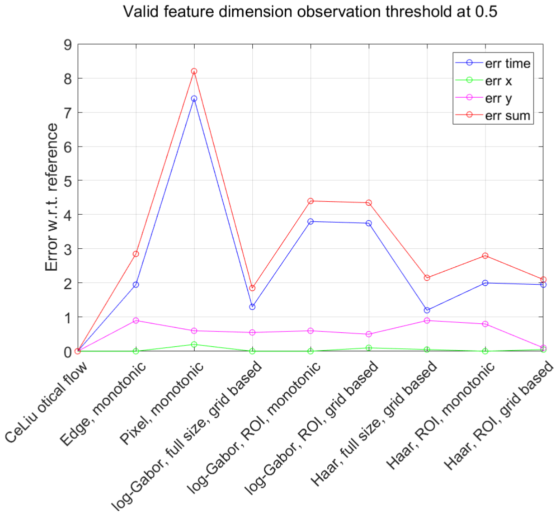

Fig. 5. Temporal block matching results compared to a reference state-of-the-art optical flow algorithm [11]. The x axis indicates the comparative approaches; feature spaces include edge, pixel, Haar-like or log-Gabor. Algorithms are the monotonic decent constraint or grid-based filter. Feature extraction is carried out on either whole images or ROI. The error is the average of 200 motion estimations using the different feature spaces and algorithms. The 0.5 in the figure title refers to an acceptance rate that at least 50% of the features must contribute to a minima.

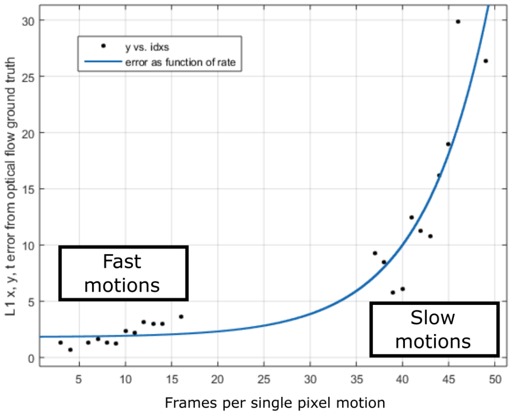

Fig. 6. Motion estimation error over time using the grid based filter approach with 10 Haar-like features applied to ROIs. The error is measured with respect to the output of [11]. As time passes, reliability of the online estimation reduces. From Figure 5 it can be seen that most of the error is introduced by ambiguity in the temporal element of matching.

VII. Discussion

The monotonically constrained version of the proposed temporal block matching approach is designed to be efficient to compute and require minimal amounts of memory. When matching the current block to a reference block, eight, 1 pixel offsets are considered; if the error of matching decreases in any particular direction, that direction can be considered as a possible direction of motion. An obvious optimisation that can

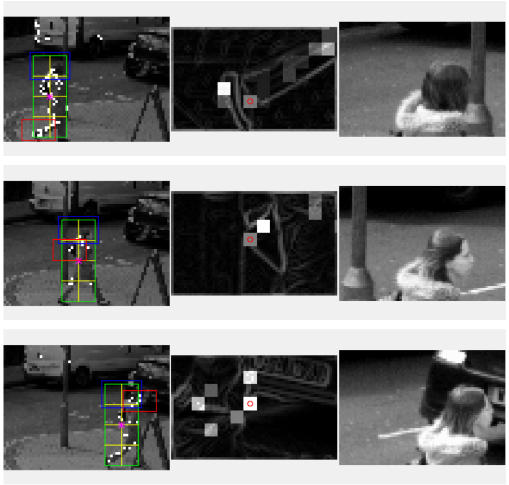

Fig. 7. The red rectangle on in the left is the ROI that currently contains the largest magnitude of differences. A random difference from the top five nearest differences to the center of the ROI is selected to track (center). The position of the head cut-outs (right) drifts as the persons horizontal motion oscillates slightly.

be implemented is; any direction that has caused the error to increase is no longer considered in the block matching search. The search will become faster as directions are eliminated until the minima in the final direction of space-time coherence is reached. Additionally in Figure 5 it can be seen that the edge based monotonic decent is almost as accurate as that computed in the 10 dimensional Haar-like space. The edge based search could be used to determine the minima in the directional matching with the Haar-like approach being applied only to the final iteration to generate an appearance model.

In section IV-A temporal block match tracking is introduced in the context of optimal grid based filtering in higher dimensional spaces. As the filter tracks through the spatio-temporal difference space, directions of motion with non-decreasing or too noisy feature responses get filtered out. When this process converges it is likely that the current temporal block has spatially moved one pixel from the reference block. At this point we have a more than 50% confidence that the motion *and* appearance of the local area-of-interest is coherent. This can be considered a form of weak learning that is not dissimilar to the approach of [16]; however the proposed approach is locally unsupervised. Rather than track the appearance of a whole object using a simple motion model, as in [16], the proposed approach can be extended to track multiple low level appearance instances with multiple simple motions without requiring significant increases in computation.

Finally, the output of the higher dimensional feature based motion estimation could be used as input to fully convolutional neural network learners. Most pixels in an image are redundant; the proposed approach extracts sparse, locally coherent, appearance models that have translated 1 pixel. The appearance models are robust to noise and changes in illumination and could be considered a source of low level ground truth. Rather than have a network attempt to learn encodings of the sparse set of coherent moving visual patterns, the output of the local models could be used as input to a network's early layers. The low resolution images used to train networks could be replaced with the filtered results of higher resolution selective imaging.

VIII. Conclusions

In this paper we describe a novel close-to-sensor computational imaging system. The architecture enables low latency ROI processing and sensor control along with concurrent interfaces for motion sensing and motor control. Low latency feedback from sub-sampled ROI processing is used to efficiently discard redundant pixels reducing the cost of processing and data transmission. High ROI sample rates minimise observed frame-to-frame difference and non-linear space-time appearance variation. This in turn enables a temporal block matching approach that provides a robust sparse motion estimation with a measure of confidence. A monotonicity constraint is shown to compare well with optimal filters and offers a high degree of efficiency. Finally the output of the sparse motion estimation can be used to build component models of complex object motion with the potential of acting as weakly supervise autonomous ground truth for higher level learning algorithms.

References

[1] B. A. Moore *et al*, "A novel method for comparative analysis of retinal specialization traits from topographic maps." *Journal of Vision*, vol. 12, 2012.

[2] D. Ballard, "Animate vision," *Artificial Intelligence*, 1991.

[3] V. Brajovic and T. Kanade, "Sensory attention: Computational sensor paradigm for lowlatency adaptive vision," in *DARPA Image Understanding Workshop*, 1997.

[4] P. Lichtsteiner, C. Posch and T. Delbruck, "An 128x128 120db 15us-latency temporal contrast vision sensor," *IEEE Journal Solid State Circuits*, 2007.

[5] S. J. Carey, D. R.W. Barr and P. Dudek, "Low power high-performance smart camera system based on scamp vision sensor," *Journal of Systems Architecture*, vol. 59, no. 10, pp. 889–899, 2013.

[6] D. Gibson, H. Muller, N. Campbell and D. Bull, "Adaptive sampling for low latency vision processing," *Asian Conference on Computer Vision, Workshop on Computational Imaging*, 2012.

[7] D. Gibson, H. Muller and N. Campbell, "Adaptive image sensor sampling for limited memory motion detection," *Pervasive and Embedded Computing and Communication Systems*, 2012.

[8] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proceedings of Imaging Understanding Workshop*, 1981.

[9] M. V. Srinivasan, "An image-interpolation technique for the computation of optic flow and egomotion," *Biological Cybernetics*, vol. 71, 1994.

[10] J. Shi and C. Tomasi, "Good features to track," *International Conference on Computer Vision and Pattern Recognition*, pp. 593–600, 1994.

[11] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, 2009.

[12] J. Snchez, E. Meinhardt-Llopis, and G. Facciolo, "Tv-l1 optical flow estimation," *Image Processing Online*, 2012.

[13] P. K. M. Tao, J. Bai and S. Paris, "Simpleflow: A non-iterative, sublinear optical flow algorithm," *Eurographics*, vol. 31, no. 2, 2012.

[14] N. G. M. S. Arulampalam, S. Maskell and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEE Transactions on Signal Processing*, vol. 50, no. 2, 2002.

[15] P. Kovesi, "Edges are not just steps," *The Fifth Asian Conference on Computer Vision*, 2002.

[16] B. Babenko, M-H. Yang and S. Belongie, "Visual tracking with online multiple instance learning," *International Conference on Computer Vision and Pattern Recognition*, 2009.