Exploring Causal Relationships in Visual Object Tracking

Karel Lebeda     Simon Hadfield     Richard Bowden
University of Surrey, Guildford, GU2 7XH, United Kingdom
{K.Lebeda,S.Hadfield,R.Bowden}@Surrey.ac.uk

Abstract

Causal relationships can often be found in visual object tracking between the motions of the camera and that of the tracked object. This object motion may be an effect of the camera motion, e.g., an unsteady handheld camera. But it may also be the cause, e.g., the cameraman framing the object. In this paper we explore these relationships, and provide statistical tools to detect and quantify them; these are based on transfer entropy and stem from information theory. The relationships are then exploited to make predictions about the object location. The approach is shown to be an excellent measure for describing such relationships. On the VOT2013 dataset the prediction accuracy is increased by 62% over the best non-causal predictor. We show that the location predictions are robust to camera shake and sudden motion, which is invaluable for any tracking algorithm and demonstrate this by applying causal prediction to two state-of-the-art trackers. Both of them benefit, Struck gaining a 7% accuracy and 22% robustness increase on the VTB1.1 benchmark, becoming the new state-of-the-art.

1. Introduction

Causality is a relation between two events, a cause (source) and an effect (consequence). In general terms, we say that an event causes another event (its effect), if it precedes the effect in time and it increases the probability of the effect happening. Although causality has been studied by philosophers for millennia, it received little attention from the scientific community before the twentieth century. Recently, theoretical advances have brought practical progress in the analysis of time series in many scientific areas.

An example of a causal relationship, which can be observed (and exploited) in the area of computer vision, is the relationship between the motions of the camera and an object in Visual Object Tracking (VOT). There are different possible causal relationships. For instance, the motion of the camera instantly causes motion of the object in the image frame. An abrupt movement of the camera (e.g., a shake) can cause a tracker to fail even in otherwise simple tracking scenarios. A particular example can be seen in the performance of all submitted trackers on the BICYCLE sequence in the ICCV VOT Challenge 2013 [25]. While this sequence is relatively easy to track in general, there are two challenging moments (see Figure 1, showing the numbers of failed trackers in the VOT Challenge). Many tracking failures are present around frame 180, caused by a strong occlusion, and around frame 140, stemming from an abrupt camera shake. If these were detected and accounted for, many of the failures could be prevented, regardless of the tracker.

Another interesting causal relationship often arises when the motion of the object causes changes of the camera motion. If there is a human in the loop, e.g., a cameraman, they are partially tracking an object by definition. A similar conclusion would hold for an automatically-controlled camera tracking the object. When the object moves towards the edge of the image, the cameraman is likely to move the camera such that the object does not disappear from the scene. An extreme case of this is the satirical Zero-order Tracker [27], shown to successfully track a challenging sequence by simply returning a bounding box on a constant location in the image. As illustrated in Figure 2, here the cameraman kept the diver in the centre of the image frame for almost whole sequence. However, even in less extreme
cases, the commonly assumed centre bias can be detected, measured and exploited.

It should be emphasised, that our work does not assume any kind of high-level oracle (e.g. a human operator) driving the camera motion. In cases where a relationship exists, we can discover and measure its influence. However, if none is present, no causal relationship is found and no further action (such as object motion prediction) is performed. This even extends to changes in behaviour within a sequence. This means the approach can be applied to any existing tracking framework to improve results. To demonstrate this, we look at its effect on two state-of-the-art trackers FoT [28] and Struck [17] on two benchmark datasets. It should be noted that while the prediction helps the tracker, it does not replace it and the result is thus still limited by the abilities of the tracker.

The contributions of this paper are as follows. After describing previous uses of causality and areas of computer vision where the causal relationships are likely to be helpful in Section 2, we show how to measure causality in an information-theoretic manner (which to our knowledge has not been used in any previous work in computer vision) and how to find the properties of the relationships found: if there is a causal relationship, what type (i.e. translation, scale,...) and direction (cause vs. effect), and what is the time delay (Section 3). Section 4 shows how these relationships can be used to give prior information to trackers processing the sequence, using two different prediction techniques. This is then experimentally evaluated in Section 5. Finally, Section 6 summarises our contributions and findings.

2. Related work

There have been numerous philosophical publications on causality in both ancient and modern times, originating from Aristotle [1] and significantly influenced by Hume [20]. We take the liberty to omit more recent philosophical publications here, for a recent overview see [30].

One of the early uses of causality for time series analysis was done by Granger [13]. He proposed a statistical causality test, determining the presence of causal relationships between two normally distributed time-series. This approach has become known as Granger causality and has been successfully used in economics [18, 35], neurosciences [11, 16], and recently in computer vision [31, 39]. Although it has been revised and improved over the decades (e.g. Hacker and Hatemi-J [15] avoided the assumption of a normal distribution), Granger causality is suitable only for linear signals, since it is based on linear regression.

More recently, a novel concept of measuring causality has been proposed: transfer entropy (TE), by Schreiber [33]. TE has since found its place in many areas, including again neurosciences [36], chemistry [3] and others [23]. To our knowledge, it has not been previously used in computer vision. As the name suggests, it is based on information theory and is therefore able to detect arbitrary non-linear relationships. In this work, we use TE to measure causation, capturing possibly complex relationships between the motion of the camera and the object.

As previously noted, causal relationships have been examined in the area of computer vision as well. Fan et al. [9] used Granger causality to explore actions and temporal dependencies between them in a surveillance scenario. This is then used to cluster and classify video-sequences, according to the actions present. In a similar direction is work learning causal relationships between events in video-sequences, which has a potential in action recognition and related tasks. An example is Fire and Zhu [10], who use Causal And-Or Graphs and Bayesian grammar models for inference about hidden effects, otherwise undetected, or Sumioka et al. [34], using causality to learn joint attention for robots. The work of Brand [4] explores the causal physics of the scene (how mechanics of objects influence other objects).

Prabhakar et al. [31] use Granger causality on sequences of keywords directly for the task of human action recognition. Yi and Pavlovic [39] perform the same task, but based on motion-capture data. They use Granger causality to infer the edges in a joint-influence graph of the human body, which improves the performance compared to fixed graphs. Finally, Narayan and Ramakrishnan [29] remove the need for motion-capture systems, using causal relationships between clusters of dense trajectories.

However, to our knowledge there has been no previous work in the field of computer vision exploiting the modern TE approach to causality estimation, and no use of causality for visual object tracking, which is the main application domain for this paper. Learned causal relationships between the motion of the camera and of the tracked object can significantly help a tracker not only to improve accuracy, but to support it during challenging events in the scene. If we were able to estimate a distribution for the object position in the current frame, based on the trajectory of the camera, this
could be supplied to the tracker as prior information. An obvious scenario is tracking by detection [2, 12, 17], a popular method for visual object tracking. The tracking is treated as a classification task, where image patches are sampled around the previous location and classified into one of the object/background classes. This can be formulated as a task of maximising the posterior probability of the object pose, where the classification score is a likelihood (given appearance). A prior probability, given by a causality-based prediction, is a very natural complement to this formulation.

On the other side of the spectrum of trackers are coupled-layer trackers [5, 26, 38], composed usually from a lower layer of independently tracked features (tracklets) and a higher layer, modelling object shape, motion, etc. The higher layer manages creation of new features such that they are likely to lay on the object, using a soft segmentation mask. This mask can be again enhanced, using prior information about the object motion, provided by a causality-based prediction.

3. Measuring the causal relationships

As mentioned previously, we use transfer entropy as a measure of causality between the camera and the object motion, a formulation employing (differential) entropies (see Sections 3.1 and 3.2). Then we employ a statistical significance analysis (Section 3.3) to discover if the relationship is significant. This is executed in each frame, until such a relationship is found. In the case it is not, we conclude that the motions are unrelated (static camera or independent motion) and we supply no information to the tracker (uniform prior), possibly until the end of the sequence. In the case where a statistically significant causal relationship is found, its parameters are estimated (Section 3.4). We then use this information to predict future object motion (Section 4) and we supply this information to the tracker.

It should be noted that it is impossible to reason about true causality without higher, semantic understanding of the scene. Therefore we work with predictive causality instead, which reasons about apparent causal links instead of true causation.

3.1. Differential Entropy

Histogram-based methods are usually employed to estimate the entropy of a random process [6, 7]. However, in our scenario this has two major disadvantages. Firstly, there is an arbitrary choice of bin size for the histograms (for quantisation of continuous signals). Secondly, the number of bins grows exponentially with the number of dimensions. This causes the histograms to be very sparse (and thus not representative of the distribution) and furthermore it requires immense computational cost even for a small number of bins per dimension.

Instead, we use differential entropy, which operates directly on the continuous variables (see Figure 3 illustrating the advantages of differential entropy). In this work we use the Kernel Density Estimation (KDE, [19]) approach to compute differential entropy, which only requires a choice of kernel (we use a Gaussian kernel with full covariance). The differential entropy of a continuous random process $X$ is

$$ H(X) = - \int_X p(x) \log p(x) \, dx, \quad (1) $$

similar to its discrete counterpart. For a finite sample set $S$ it is approximated using KDE by:

$$ \hat{H}(X) = - \frac{1}{|S|} \sum_{x_i \in S} \log \hat{p}(x_i) $$

$$ = - \frac{1}{|S|} \sum_{x_i \in S} \log \left( \frac{1}{|S|} \sum_{x_j \in S} \kappa_{xy}(x_i - x_j) \right), \quad (2) $$

where $\kappa_{xy}$ is a Gaussian kernel with covariance $\Sigma$ (estimated from the data using expectation maximisation). The probability $p(x_i)$ outside the logarithm is approximated by the distribution of the samples from $S$ (i.e. assuming $S$ was drawn according to $p(x)$).

3.2. Transfer entropy

Transfer entropy is a measure of directed influence flow between two processes ($X \rightarrow Y$, with windows$^1$ of length $n$ and lag $\Delta t$). For continuous signals we define it as:

$$ T_{X \rightarrow Y} = \iint \log \frac{p(y_{t+n}, x_{t+n-\Delta t}^n | y_{t-n}^n)}{p(y_{t+n} | x_{t-n}^n)} \, dy_t \, dy_{t-n} \, dx_{t-n-\Delta t}^n, \quad (3) $$

with time windows defined as $y_{t-n}^n = (y_{t-n}, \ldots, y_{t-1})$. It can be reformulated (using differential entropies) as the differen-

$^1$The window lengths do not necessarily need to be equal for $X$ and $Y$. 

![Figure 3. Comparison of discrete and differential entropy. Points were uniformly sampled from an interval [0,1]. The histogram bins for discrete entropy computation were fixed at integer positions. Notice how stable differential entropy is, even with sparsely distributed points (no interpolation used).](image-url)
ence of two information gains:

$$T_{X \rightarrow Y} = (H(Y_t, \bar{Y}_t^n) - H(Y_t^n)) - (H(Y_t, \bar{Y}_t^n, X_{t-\Delta t}^n) - H(Y_t^n, X_{t-\Delta t}^n)) .$$

(4)

Intuitively, this tells us that if there is a causal relationship between $X$ and $Y$ with the correct direction and lag, then adding knowledge about $Y_t$ brings more information to a system which does not know $X$, than to one which does (as $X$ can partially predict it).

### 3.3. Statistical significance analysis

Once we know the transfer entropy between the motion of the camera and the object, we need to decide if this relationship is significant enough to make predictions of the object movement. Tests of statistical significance are preferred rather than comparing to a fixed threshold, since they offer theoretically founded decisions with probabilistic thresholds and explicitly cope with the inherent uncertainty caused by insufficient data.

To provide a sequence-specific baseline with no causal relationship the target time series $Y$ is shuffled to remove any causality while retaining the distribution of amplitudes. We denote the shuffled signal as $\bar{Y}$. Then a Welch’s t-test is performed, to obtain a p-value indicating the probability that both $T_{X \rightarrow Y}$ and $T_{X \rightarrow \bar{Y}}$ arise from the same distribution (a null hypothesis). When the observed causal relationship is statistically more significant than what is likely to arise by chance given the signal distributions, we conclude that it can be used for object motion prediction. This approach is shown to be successful in 15 out of 16 sequences from the VOT2013 dataset.

### 3.4. Finding the optimal parameters

When the causal relationship has been confirmed, we can attempt to predict the object motion from the overall movement of the whole scene (dual to the camera motion). Since different processes have, in general, different causal relationships, each particular sequence will have unique properties. In other words, we need to find an optimal set of parameters for subsequent prediction. These parameters are the time delay and the length of the time window, which can be seen as a mean and variance of the lag $\Delta t$. We want to pick these such that TE is maximal. Unfortunately, for the window length $n$ this may not be as simple as in the case of $\Delta t$, as the transfer entropy stays high even when the window length is over-estimated (see Figure 4). We want to ensure that we do not miss any important information while using only important information. Excessively long time windows make the prediction unnecessarily slow without adding any significant gain. Also, non-discriminative features are likely to degrade performance, particularly with small training sets [14]. Therefore we define a relative improvement measure $f(\Delta t, n)$ (visualised by the column colours in Figure 4) from adding an additional frame to the window length and we require this relative improvement to be higher than a given threshold. The maximisation is then constrained as follows:

$$(\Delta t^*, n^*) = \arg \max_{\Delta t, n} T(\Delta t, n) \quad \text{s. t.} \quad f(\Delta t, n) > \theta_f ,$$

(5)

with $f$ defined as

$$f(\Delta t, n) = \frac{T(\Delta t, n) - \max_{\Delta t, n < n} T(\Delta t, n)}{\max_{\Delta t, n \leq n} T(\Delta t, n)} ,$$

(6)

where $T(\Delta t, n)$ relates to TE parameterised with a particular window length and lag. For experiments in this publication, $\theta_f = 10\%$ was used.

Figure 4 shows the effect of the parameters $\Delta t$ and $n$. Notice the characteristic triangular shape of the area with consistently high TE: when we extend the time window, already containing the most relevant information, no significant information is gained or lost.

### 4. Predicting the object motion

When using the video data, the signals are defined as follows. We assume having two multivariate time-series, $\mathbf{I}$ for the camera (image) and $\mathbf{O}$ for the object. We use multivariate signals: $x$ and $y$ coordinates and size (bounding box diagonal length), but additional dimensions (such as rotation) are possible. For the camera, the measured quantity is the image position relative to the first image. This is expressed in pixels, and is defined by the accumulated inter-frame motion $(\Delta x_t, \Delta y_t, \Delta s_t)^\top$ i.e. $\mathbf{I}_t = (\sum_{r=1}^t \Delta x_r, \sum_{r=1}^t \Delta y_r, s_0 \prod_{r=1}^t \frac{\Delta s_r}{\Delta s_1})^\top$.

The global motion of the camera can be estimated robustly using the inter-frame shift of the whole image, with higher reliability than the object tracking. In our implementation we use a simple approach based on feature matching and RANSAC, but a more complicated method (e.g.
based on tracklets like in FoT [28]) may be used in challenging scenarios. Therefore, any discovered relationship can be used to transfer information from one (reliably estimated) signal to the other. In this publication, this information transfer is seen as the estimation of a distribution of possible poses for an object, based on its history $O$ and additionally on its relation to the information from the image signal $I$. This distribution can be supplied to a tracker as prior information to guide the tracking process. We examine two different approaches to object position prediction; the following sections describe these. For an intuitive comparison of both methods of prediction see Figures 5 and 6 (only $x$ coordinate prediction is shown).

4.1. Window-based prediction

In the first case, a window-based prediction is used, similar to a non-linear autoregressive model. This approach is intuitively closer to the transfer-entropy background as described in Section 3.2. In autoregression, the current state is estimated (predicted) using a learned autoregressive function $\phi_w$ from its own history: $y_t = \phi_w(y_{t-n}^n)$. Knowing there is a causal relationship between the two signals, the prediction can be improved using the other signal: $y_t = \phi_w(y_{t-n}^n, x_{t-\Delta t}^\ast)$, or more particularly:

$$O_t = \phi_w(O_{t-n}^n, I_{t-\Delta t}^\ast).$$ (7)

The window-based regressive function $\phi_w$ can be learned, taking a machine-learning approach. In other words, we take a set of all windows from the history and learn a regression (mapping) from the known part ($O_t^n$ or both ($O_t^n, I_{t-\Delta t}^\ast$) knowing the optimal parameters) to the current pose $O_t$. The principle of this is visualised in Figure 5.

4.2. Time-based prediction

In the second case, both the object position and the image position are modelled as functions of time $I_t$ and $O_t$. A sequential version of the autoregressive function can be learned, using the information about the data sequentiality: $O_t = \phi_t(t \mid O_{1...t-1})$.

Exploiting the causal knowledge, the $I$ signal is shifted forward by the lag found as described in Section 3.4, Equation (5), to create $I_{t-\Delta t}^\ast$ (aligning the signals). Then machine learning is used again, to learn the relationship between the two time-aligned signals, and to predict the future changes of $O$. We again define a (time-based) regressive function $\phi_t$ such that

$$O_t = \phi_t(t \mid O_{1...t-1}, I_{1...t-\Delta t}^\ast, n^*).$$ (8)

The window length $n^*$ is used as a measure of uncertainty in the timing of $I$, i.e. how large a part of $I$ is necessary to be taken into account during the prediction. In other words, both $I$ and $O$ are modelled as functions of time with the former guiding prediction of the latter in areas of insufficient data. The time-based function $\phi_t$ is visualised in Figure 6.

5. Experimental evaluation

For our experiments, we implemented the proposed method as follows. On the signals $I$ and $O$ we perform the causality analysis, using $T_{I \rightarrow O}$; the $I$ signal takes the role of $X$ as used in Section 3.2 while $O$ represents $Y$. For an initial coarse estimation of the signals lag, several overlapping windows with fixed length are used and TE with its statistical significance is computed for each of them in each frame, using Equation (4). When the statistical significance of any window exceeds a specified significance level $\alpha$, a causal relationship is assumed and the optimal set of parameters found according to Equation (5). If no window is significant enough, we assume there is currently no causal relationship between the camera and the object motion. In our experiments, we used $n = 4$ and $\Delta t \in \{-4, -7, -10, \ldots\}$, and a conservative significance level $\alpha = 0.01 \%$.

The prediction was carried out as described in Sections 4.1 and 4.2. For the machine-learning stage, Gaussian Process Regression (GPR) was employed [32], as a probabilistic non-parametric regression approach, robust to overfitting. In all cases we used a combination of an RBF
and a bias kernel. Since different video-sequences in general do not share their causal properties, all the predictions were made using online sequence-specific learning. In other words, a tracker is required to track successfully for some time at the beginning of the sequence and this initialisation is used to learn the properties of the sequence (this implies that early tracking failures may lead to incorrect causal relationships, which would be however rejected as not significant). The prediction would be then supplied to the tracker as prior information and its tracking result would be added to the history data for a new prediction. For time efficiency, the φ-functions learning was initialised using φ from the previous frame.

For the window-based autoregression, windows containing a short history (3 frames: O1) of the position signal O were taken as features to predict the position in the consecutive frame O2 (to learn φa). Any other temporal information (inter-window relationships) were discarded, treating the data as a bag of equally important training inputs. The function φa was then learned and queried with the current history window O2 to obtain the prediction. The window-based causal prediction was done in a similar manner. The history windows O1 and I1−Δt were concatenated together into (3 × 2 × n) -dimensional features (3× because of both I and O being (x, y, s) vectors), and the function φw was trained on the available history.

In the case of the time-based prediction via φx, the independent features are simply the frame indices and the dependent features the coordinates. For the causal prediction (the φ function), we need a technique to tie two signals together in an a priori unknown relationship. This can be achieved using a coregionalisation in the GPR. Coregionalisation is a technique which can model both signals O and I as functions of time with a hidden relationship. Knowing the shape of one of the signals (I) then guides the prediction of the other one (O) even in locations distant from any training points of O, as shown in Figure 6.

There are periods in the sequences, where no causal relationship was detected, and therefore no prediction parameters exist. In such places, the causal prediction is replaced by the appropriate autoregressive function: φw by φa and φt by φs. This explains the identical results for the sequences without causality during qualitative experiments in Section 5.3 (HAND, JUMP and TORUS).

5.1. Evaluation of causality detection

For the causality detection evaluation, the ICCV VOT Challenge 2013 [25] dataset was used (16 sequences, each containing between 172 and 770 frames). There are no “ground-truth causal relationships” we could use to measure the quality of our detection on the sequences (with the exception of zero relationship in case of static camera). However, the detected relationships are consistent with intuitive understanding of the scene dynamics and the optimal prediction parameters fulfilled our expectations. This shows that it is possible to use information-theory based measures to discover and quantify relationships between signals in real sequences for the task of visual object tracking. Using these, we can measure if there is a causal relationship between a camera and an object in a video-sequence, in which parts of the sequence, and we can measure its properties.

See Table 1 for the results. In the third column, we show the fraction of the sequence marked as containing a significant causal relationship. Then the optimal parameters for prediction are shown; in the case of different relationships for different time periods there are multiple parameter sets (e.g. 1 for BICYCLE, none for HAND or 3 for SINGER).

In the case of the sequence JUMP, none of the detected causal relationships were statistically significant. However, this is not necessarily an error as although the camera is not static, we do not know if a true relationship exists between object and camera motion. There are three sequences with a static camera (constant zero I) and therefore no causal relationships, these are marked in grey in the tables. For two of them, this was correctly detected using TE. For the CAR sequence, a causal relationship was incorrectly discovered due to inaccurate estimation of I. However, this means that causality detection only failed in 1 out of 16 sequences. It is also worth pointing out the relatively common occurrence of the (-1,1) pair, indicating the immediate causal effect of a moving camera on the apparent motion of a static object.

See the supplementary material for causality detection evaluation on synthetic data, where the GT is known.

5.2. Qualitative prediction evaluation

The task we are given during the prediction stage is to estimate the distribution of the possible object positions to be supplied to the tracker as a prior. For a performance measure, there is a requirement to discover how well the ground truth position (GT) is represented by this distribution. While simple distance between the distribution mean

| Sequence | Length | Length ratio | Δt | n*
|----------|--------|--------------|----|--
| BICYCLE  | 221    | 82.5%        | -3 | 7
| BOLT     | 350    | 64.3%        | -1 | 2
| CAR      | 374    | 64.7%        | -10,-14 | 7.5
| CUP      | 303    | 53.5%        | -3, -3 | 4.8
| DAVID    | 770    | 96.9%        | -2,-1 | 2.1
| DIVING   | 231    | 40.3%        | -11 | 8
| FACE     | 415    | 91.3%        | -1 | 1
| GYMNASTICS | 207 | 81.6%        | -1 | 1
| HAND     | 244    | 60.0%        | NA  | NA
| ICESKATER| 500    | 92.4%        | -11,-2 | 3.7
| JUICE    | 404    | 90.6%        | -1 | 1
| JUMP     | 228    | 0.0%         | NA  | NA
| SINGER   | 351    | 82.3%        | -17,-13,-11 | 6.1,1
| SUNSHADE | 172    | 59.3%        | -8 | 8
| TORUS    | 264    | 0.0%         | NA  | NA
| WOMAN    | 597    | 49.6%        | -8,-3,-3 | 5,8,1
| Average  | 355    | 59.1%        |     |     

Table 1. Causal detections on the VOT2013 dataset – detected durations and properties.
and the GT indicates the prediction accuracy, it does not take uncertainty into account. In particular, if two predictors predict the same correct position, the one with high confidence is of most benefit to the tracker. This holds in the opposite direction as well, for a misprediction it is better to report lower certainty. For these reasons, we used the probability density function (PDF) as a performance measure in our experiments. Additionally, we have integrated the error over the entire PDF support region to obtain expectation for the prediction error. In the following section, we use the mean values of both across each sequence as quantitative measures.

The causality-based prediction was compared with several alternative approaches as follows. Firstly, two naive approaches are examined, treating all historical states as equally important, based on the assumption that the object stays in an approximately stable location. One models the distribution as a Gaussian, as used in the re-detector of Lebeda et al. [26], while the other one uses a Kernel Density Estimation (KDE) to model the PDF with higher accuracy. These can be seen as implementation of the central bias and stable location priors. Visual tracking algorithms often use a Kalman filter (KF), or its extension, as their internal motion model [8, 22, 24, 40]. Therefore, the KF is a natural alternative to causal prediction. Finally, we evaluated the autoregressive functions \( \phi_a \) and \( \phi_s \), and the causal predictions \( \phi_w \) and \( \phi_t \).

The results are visualised in Figure 7. In both cases, the GT is well inside the naively predicted distributions. However, these distributions are spread over the whole image and therefore the PDF is relatively low. In the case of the KF, the predictions lag behind the true signal somewhat, causing mispredictions. The autoregression given by \( \phi_a \) and \( \phi_s \) helps significantly with the GT being at least at the edge of the predicted distribution; \( \phi_s \) shows better performance than \( \phi_a \). Window-based causal prediction gives accurate modes for the distribution, although the long window in the case of Diving (\( \eta^* = 8 \)) results in a low confidence and thus a lower PDF. The time-based prediction performed the best of all the tested predictors; \( \phi_t \) predicted positions close to the GT while having an appropriate confidence.

5.3. Quantitative prediction evaluation

As previously mentioned, the mean PDF and mean expectation of error across each sequence were used as performance measures in the quantitative evaluation. See Table 2 for the results. In the Car sequence, the tracked car stops for a large part of the sequence in one location, the KDE predicted very high probability for this location, which is reflected by a very high mean PDF. A similar phenomenon can be found in Gymnastics, with the tracked person standing in one place for a part of the sequence.

In the Bolt and IceSkater sequences, the I signal estimation failed for one region of each sequence due to very low texture of the background. This renders the relationship between the camera and object motions unstable and therefore the \( \phi_w \) and \( \phi_t \) predictors have lower performance in these sequences. This is more noticeable in the case of error expectation, where these outliers render the \( \phi_t \) prediction to not have the lowest average error, despite being the lowest on majority of sequences.

In general, disregarding these outliers, several statements can be made about the performance of the compared predictors. Both global probability distributions have an image-wide spread and therefore a very low PDF. Prediction using KF is better localised and has therefore significantly better performance, although still worse than the learned regressive functions. For the learned functions, we can say that time-based ones (\( \phi_w \) and \( \phi_t \)) in general perform better than window-based \( \phi_a \) and \( \phi_s \). Regressive function \( \phi_w \) performs slightly worse than its non-causal counterpart \( \phi_a \), due to the lower confidence of the prediction (higher variance and therefore lower PDF). The time-based causal function \( \phi_t \) was shown as the best predictor, beating the second best by a large margin (62%). In addition, it performs more than three times better than KF, which is a commonly used motion model.

Table 3 show the effect of causal prediction on the performance of the state of the art trackers FoT [28] and Struck [17]. The tabulated values are the VOT accuracy/robustness metrics [25] – mean bounding box overlap (higher is better) and number of failures per sequence (lower is better). We compare against vanilla trackers and a simple background motion compensation (BMC), using the image context but no temporal causal relationships. While the simple camera motion information does not prove useful, supplying the tracker prior information from causality-based prediction improves its performance significantly. In general, robustness is affected only slightly, while the main improvement is in the accuracy domain. For
In this paper, we have explored causal relationships between object and camera motions. We have proposed an approach to discover and quantify this relationship using transfer entropy, a statistical tool which to our knowledge has not been used in any previous publication in the area of computer vision. We have also shown that it is possible to find the optimal time window for prediction of the object position based on the global image motion even for complex non-linear relationships. Finally, these causality-based motion predictions were evaluated on a range of standard tracking sequences, and shown to offer excellent performance (increasing average prediction accuracy by 62 % and improving the top performing tracker on VTB1.1 by 7 % in accuracy and 22 % in robustness), with particular robustness to camera shakes and fast motion. These are typically the greatest source of errors in modern tracking, as shown in the recent VOT challenge, and thus the proposed techniques, which we will make publicly available, provide an invaluable addition to any tracking algorithm.

**Acknowledgement:** This work was supported by the EPSRC project EP/I011811/1: Learning to Recognise Dynamic Visual Content from Broadcast Footage, and the Rubin Ezra Scholarship.

### 6. Summary

In this paper, we have explored causal relationships between object and camera motions. We have proposed an approach to discover and quantify this relationship using transfer entropy, a statistical tool which to our knowledge has not been used in any previous publication in the area of computer vision. We have also shown that it is possible to find the optimal time window for prediction of the object position based on the global image motion even for complex non-linear relationships. Finally, these causality-based motion predictions were evaluated on a range of standard tracking sequences, and shown to offer excellent performance (increasing average prediction accuracy by 62 % and improving the top performing tracker on VTB1.1 by 7 % in accuracy and 22 % in robustness), with particular robustness to camera shakes and fast motion. These are typically the greatest source of errors in modern tracking, as shown in the recent VOT challenge, and thus the proposed techniques, which we will make publicly available, provide an invaluable addition to any tracking algorithm.

### Table 3. Tracking results on the VOT2013 benchmark. The best and second best results are highlighted separately for FoT and Struck for accuracy/robustness.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>FoT</th>
<th>Struck</th>
<th>Struckstr</th>
<th>Struckstr</th>
</tr>
</thead>
<tbody>
<tr>
<td>BICYCLE</td>
<td>0.70</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>BOLT</td>
<td>0.46</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>CAR</td>
<td>0.55</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>CUP</td>
<td>0.38</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>DAVID</td>
<td>0.76</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>DIVING</td>
<td>0.25</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>FACE</td>
<td>0.74</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>GYMNASTICS</td>
<td>0.66</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>HAND</td>
<td>0.40</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>ICESKATER</td>
<td>0.43</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>JUMP</td>
<td>0.62</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>SINGER</td>
<td>0.74</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>SUNSHADE</td>
<td>0.38</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>TURIS</td>
<td>0.38</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>WOMAN</td>
<td>0.61</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>Average</td>
<td>0.62</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
</tr>
</tbody>
</table>

### Table 4. Tracking results on VTB1.1. Results in the last three columns were obtained using the VOT evaluation criteria, using the VTB criteria would improve the accuracy even further.

<table>
<thead>
<tr>
<th>Category</th>
<th>ASLA21</th>
<th>SCM41</th>
<th>Struck</th>
<th>Struck</th>
<th>Struck</th>
<th>Struck</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>0.59</td>
<td>0.39</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>DEF</td>
<td>0.51</td>
<td>0.39</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>FM</td>
<td>0.42</td>
<td>0.39</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>IPR</td>
<td>0.52</td>
<td>0.39</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>TV</td>
<td>0.69</td>
<td>0.39</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>LR</td>
<td>0.56</td>
<td>0.39</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>MB</td>
<td>0.45</td>
<td>0.39</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td>OCC</td>
<td>0.56</td>
<td>0.39</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>OPR</td>
<td>0.56</td>
<td>0.39</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>OV</td>
<td>0.54</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>SV</td>
<td>0.54</td>
<td>0.39</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>All</td>
<td>0.53</td>
<td>0.39</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
<td>0.53</td>
</tr>
</tbody>
</table>
References


[34] H. Sumioka, Y. Yoshikawa, and M. Asada. Learning of joint attention from detecting causality based on transfer entropy. *J. of Robotics and Mechatronics*, 2008. 2


