Top-Down, Bottom-Up Multivalued Default Reasoning for Identity Maintenance

Vinay D. Shet, David Harwood, and Larry S. Davis
Computer Vision Laboratory, University of Maryland
College Park, MD, USA
{vinay,harwood,lsd}@umiacs.umd.edu

ABSTRACT

Persistent tracking systems require the capacity to track individuals by maintaining identity across visibility gaps caused by occlusion events. In traditional computer vision systems, the flow of information is typically bottom-up. The low level image processing modules take video input, perform early vision tasks such as background subtraction and object detection, and pass this information to the high level reasoning module. This paper describes the architecture of a system that uses top-down information flow to perform identity maintenance across occlusion events. This system uses the high level reasoning module to provide control feedback to the low level image processing module to perform forensic analysis of archival video and actively acquire information required to arrive at identity decisions. This functionality is in addition to traditional bottom-up reasoning about identity, employing contextual cues and appearance matching, within the multivalued default logic framework proposed in [18]. This framework, in addition to bestowing upon the system the property of nonmonotonicity, also allows for it to qualitatively encode its confidence in the identity decisions it takes.

Categories and Subject Descriptors

I.2 [**Vision and Scene Understanding**]: Perceptual Reasoning

General Terms

Algorithms

Keywords

Visual Surveillance, Nonmonotonic reasoning, Default Logic, Identity Maintenance, Tracking

VSSN'06, October 27, 2006, Santa Barbara, California, USA.

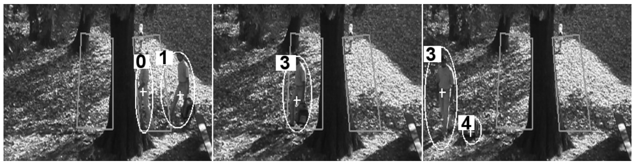

Figure 1: Sequence of images showing individual 0 and 1 disappearing from view and 3 subsequently appearing from behind an occlusion

1. INTRODUCTION

Automated visual surveillance systems require the capacity to detect and persistently track objects from their point of entry into the field of regard of the cameras, to their point of exit. This is, however, an extremely difficult task as object tracks are often lost due to occlusions by static structures in the scene or interactions with other tracked objects. Figure 1 shows individual 0 and individual 1 disappearing from view and subsequently individual 3 appearing in view from behind an occlusion. A persistent tracking system should be able to determine whether $individual(3) = individual(0)$ or $individual(3) = individual(1)$ and resume tracking. Moreover, if multiple individuals interact, then regardless of what goes on during the interaction event, when the individuals separate, the system should correctly establish identity of the individuals involved.

Traditionally, the problem of identity maintenance after occlusions has been handled by appearance matching. The basic premise is that if two individuals appear similar to each other than they must be equal, while if they appear dissimilar, then they must be not equal. However, it is additionally possible to employ context based cues to perform identity maintenance. For example, in Figure 1, individual 3 drops a bag after emerging from the occlusion. Under certain circumstances, it might be possible to conclude that $individual(3) = individual(1)$ based on the fact that both of them are carrying a similar looking bag. However, it is generally difficult to detect that either of 0 or 1 was carrying a bag. The fact that 3 was carrying a bag would be discovered first (e.g. through background subtraction) because a "bag-like" object, 4, splits from 3. Given this information, and assuming that bag 4 did not change hands during the short visibility gap, searching for the image of bag 4 in the archived images of 0 and 1 can lead us to conclude that $individual(3) = individual(1)$.

In most vision systems the flow of information is usually bottom-up. The low level computer vision routines are run first to gather information which is then provided to high level reasoning routines. However, the situation described above requires information to flow top down. The reasoning module has to understand that there exists a deficit of information (perhaps because appearance matching by itself was unable to distinguish between the individuals in Figure 1) and given that 3 was carrying a bag, try to actively search archival video for the presence of a bag in images of individuals 0 or 1. This example captures the general problem of context driven image analysis that we address here.

In [18], we described a system to perform identity maintenance across possibly large visibility gaps by employing, in addition to traditional appearance matching, several context based cues. We proposed a multivalued default logic (MVDL) framework in which these cues were regarded as different sources of information regarding the truth value of a given equality statement and were integrated in an information centric manner. This system recognized the occurrence of low level "atomic" events from the input video and provided this information to the high level MVDL reasoning framework where identity decisions were made based on low level observations. The flow of information was strictly bottom-up.

Here, we describe a system that treats occlusions and object interactions as closed world events and uses the MVDL framework to explicitly reason about object identities upon re-appearance. Specifically our contribution is the use of the high level reasoning framework to actively resolve states representing contradictions or lack of information regarding equality of individuals, by providing control feedback, and driving low level image processing modules.

2. RELATED WORK

Object occlusions have been handled in the literature in various ways. Pfinder [21] represents models of humans by a collection of colored blobs and handles partial occlusions. [7] presents a Bayesian blob-tracker which implicitly handles occlusions by incorporating the number of interacting persons into the observation model and inferring it using Bayesian Network. [8] accounts for occlusions by an outlier component in a generative appearance model and uses online EM to learn and update the parameters of this model. [6] uses closed-world regions to perform context-based tracking of multiple objects with erratic movements and collisions. [11] tracks vehicles by using a ground plane constraint to reason about vehicle occlusions. [5, 13] use region tracks and appearance models to identify people after occlusions. [17] uses appearance models to localize objects and attempts to infer object's depth ordering. [9] and [19] both model videos as a layered composition of objects and use EM to infer objects appearances and motions. [15] represents self-occlusion with layered templates, and uses a kinematic model to predict occlusions. [2] automatically decomposes video sequences into constituent layers sorted by depths by combining spatial information with temporal occlusions. Identity maintenance in surveillance has typically employed some form of appearance matching such as color and shape [14], gait [1] or face recognition [23] . Microsoft's *EasyLiving* project [12] employs two stereo cameras to track up to 3 people in a small room while [20] describes a multi-camera indoor people localization in a cluttered environment.

3. OVERVIEW

Persistent tracking systems have to contend with two kinds of events. The first kind is when a tracked object is occluded by a static scene structure such as a tree or a pillar and the second kind is when two or more tracked objects interact, visually merging into one. When multiple tracked objects interact, our system does not attempt to separate out the individual objects that make up the merged region. It, instead, maintains in its knowledge base that, for the duration of the interaction event, the region being tracked is composed of multiple objects. Henceforth, occlusions caused both by static structures in the scene as well as object interactions will be collectively referred to as occlusion events.

As mentioned earlier, in addition to appearance matching, there exist other identifying cues that can provide information about an individual's identity. Knowledge about these cues and object behavior is encoded in our system as rules in a logic programming language. Since this knowledge represents behavior of objects in a real world, it can never be definite and completely correct. We therefore employ default logic (which is the core of the MVDL framework) as the language to specify these rules, which provides our framework the important property of nonmonotonicity (the property of retracting or disbelieving old beliefs upon acquisition of new information). Default rules capture information that are "true by default" or "generally true" but may cease to be true in the future.

The MVDL framework bestows upon the system, in addition to the property of nonmonotonicity, the capacity to qualify identity decisions with a qualitative confidence measure. Identity rules included in the MVDL framework are applied to reason about object identity when an object appears from occlusion events. These rules model occlusion events as closed world spaces; meaning that an object that enters an occlusion event should (with exceptions) exit from it eventually. Default logic models such assumptions while maintaining the possibility for them to be incorrect. If only a single object is involved in an occlusion event, such as an individual walking behind a tree and later reappearing, the uncertainty involved in making an identity decision is the least (as compared to scenarios where multiple objects are involved) and therefore, such a default rule will have a very high priority (or certainty) in the MVDL framework. This is in contrast to a situation where multiple individuals are involved in an occlusion event and the system is forced to establish identity solely on the basis of appearance matching. Identity decisions made on the basis of solely appearance matching typically have the lowest priority.

Unfortunately it is often the case that the only information the system possesses is the appearance matching score. In such cases, the system might quickly enter a state of no information or contradiction regarding identity of a few individuals. To emerge from this state, the system actively searches for specific objects of interest that will make identification more certain. The system is composed of two layers, with the low level module performing object detection, local low level tracking and fact generation, and the high level module responsible for identity maintenance and providing control feedback to the low level module for conducting searches focused on archival video.

4. REASONING FRAMEWORK

Logic programming systems employ formulae that are either facts or rules to arrive at inferences. In visual surveillance, rules can be used to define various activities of interest as well as make inferences about identity of objects. Rules are of the form "$A \leftarrow A_0, A_1, \cdots, A_m$" where each A_i is called an atom and ',' represents logical conjunction. Each atom is of the form $p(t_1, t_2, \cdots, t_n)$, where t_i is a term, and p is a predicate symbol of arity n. Terms could either be variables (denoted by upper case alphabets) or constant symbols (denoted by lower case alphabets). The left hand side of the rule is referred to as the head and the right hand side is the body. In first order logic, rules of this type are interpreted as "if body then head". Facts are logical rules of the form "$A \leftarrow$" (henceforth denoted by just "A") and correspond to the input to the inference process. These facts are the output of the low level computer vision algorithms, and include "atomic" events detected in video (entering/exiting the scene, dropping of a package) and data from background subtraction and tracking. Finally, '$\neg$' represents negation such that $A = \neg\neg A$.

4.1 Multivalued Default Logic

Traditional first order logic rules express that if their body evaluates to true then the head always evaluates to true. The size of the inference set in such logics always monotonically increases as new facts are inserted into the knowledge base. Default logic [16], on the other hand, expresses rules that are "true by default" or "generally true" but could be proven false upon acquisition of new information in the future. This property of default logic, where the truth value of a proposition can change if new information is added to the system, is called nonmonotonicity.

Definition 1 (Default Theory). *A default theory Δ is of the form $\langle W, D\rangle$, where W is a set of traditional first order logical formulae (rules and facts) also known as the definite rules and D is a set of default rules of the form $\frac{\alpha:\beta}{\gamma}$, where α is known as the precondition, β is known as the justification and γ is known as the inference or conclusion.*

A default rule of this form expresses that if the precondition α is known to be true, and the justification β is consistent with what is currently in the knowledge base, then it is possible to conclude γ. Such a rule can be also written as $\gamma \leftarrow \alpha, not(\neg\beta)$. 'not' represents the negation by "failure to prove" operator and the consistency check for β is done by failure to prove its negation.

"If two objects 'a' and 'b' appear similar to each other, then generally they are one and the same" is a default rule. A rule such as this would be specified as

$$\begin{aligned} equal(P_1, P_2) \quad \leftarrow \quad & appear\_similar(P_1, P_2), \\ & not(\neg equal(P_1, P_2)) \end{aligned}$$

This rule will continue to infer that two objects are equal as long they appear similar (using some appearance matching algorithm) and it is consistent to believe that the two objects are equal i.e. until the point when it can derive $\neg equal(P_1, P_2)$ via some other rule.

Default theories typically have multiple rules for each proposition representing both positive and negative inferences. For example, a default theory that reasons about identity would typically have multiple rules for inferring *equal* and $\neg$*equal*, each rule representing a different line of reasoning. In visual surveillance, various rules in the system are regarded as different sources of information concerning the truth value of a given proposition. It is important to note that by truth value we mean our degree of belief in the veracity or falsity of a given proposition. This is different from the actual truth value of the proposition in the real world. These sources contribute different amounts of information to the decision making process and consequently our degree of belief in these propositions should mirror the information content. For example, default rules are not always correct and could be proven wrong by definite rules. Therefore, in this approach, definite rules provide more information than default rules. Also, among default rules themselves, it is possible that some rules provide more information than others. This suggests a prioritization of different rules in the system in order of their information content.

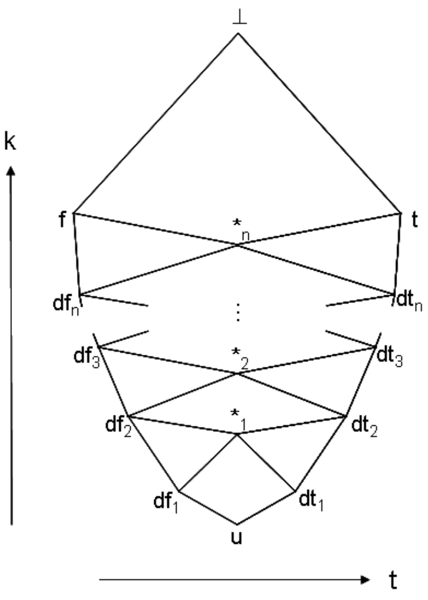

Figure 2: Bilattice for prioritized default logic.

4.2 Bilattice

Bilattices [4] provide an elegant and convenient formal framework in which the information content from different sources can be viewed in a truth functional manner. Truth values assigned to a given proposition are taken from a set structured as a bilattice. Informally a bilattice is a set, B, of truth values composed of two lattices $(B, \wedge, \vee)$ and $(B, \cdot, +)$ each of which is associated with a partial order $\leq_t$ and $\leq_k$ respectively. $\wedge$ and $\cdot$ correspond to the greatest lower bound (glb) operators while $\vee$ and $+$ correspond to the lowest upper bound (lub) operator. The $\leq_t$ partial order indicates how true or false a particular value is, with f being the minimal and t being the maximal. The $\leq_k$ partial order indicates how much is known about a particular sentence. The minimal element here is u (completely unknown) while the maximal element is $\perp$ (representing a contradictory state of knowledge where a sentence is both true and false). The glb and the lub operators on the $\leq_t$ partial order ($\wedge$ and $\vee$) correspond to the usual logical notions of conjunction and distinction, respectively. The lub operator on the $\leq_k$ partial order, $+$, corresponds to the combination of evidence from different sources or lines of reasoning while the glb operator on the $\leq_k$ partial order, $\cdot$, corresponds to the consensus operator. A bilattice is also equipped with a negation operator $\neg$ that inverts the sense of the $\leq_t$ partial order while leaving the $\leq_k$ partial order intact.

Figure 2 shows a bilattice corresponding to prioritized default logic. The set B of truth values contains, in addition to the usual definite truth values of t and f, dt_i and df_i corresponding to true-by-default (also called "decided-true") and false-by-default (also called "decided-false"), u corresponding to "unknown", $*_i$ corresponding to "undecided" (indicating contradiction between dt_i and df_i),$(i = 1 \cdots n)$ and $\perp$ corresponding to "contradiction" (between t and f). The t-axis reflects the partial ordering on the truth values while the k-axis reflects that over the information content. This bilattice provides a correlation between the amount of information and the degree of belief in a source's output. Obtaining more information about a proposition, indicated by rising up along the k-axis, causes us to move away from the center of the t-axis towards more definitive truth values. The only exception to this being a contradiction, in which case, we move back to the center of the t-axis. Negation corresponds to reflection of the bilattice about the $\perp -u$ axis. It is also important to note the this bilattice is distributive with respect to each of the four operators.

4.3 Inference

Based on this framework, it is possible to define truth tables for each of the four glb and lub operators. Inference in such systems is defined in terms of closure. If $\mathcal{K}$ is the knowledge base and ϕ is a truth assignment labelling each sentence $k \in \mathcal{K}$ with a truth value then the closure of ϕ, denoted $cl(\phi)$, is the truth assignment that labels information entailed by $\mathcal{K}$. For example, if ϕ labels sentences $\{p, q \leftarrow p\} \in \mathcal{K}$ as true; i.e. $\phi(p) = T$ and $\phi(q \leftarrow p) = T$, then $cl(\phi)$ should also label q as true as it is information entailed by $\mathcal{K}$. Entailment is denoted by the symbol '$\models$' ($\mathcal{K} \models q$). If S and S' are sets of sentences such that $S \models q$ and $S' \models \neg q$, then the truth value assigned to q is given by

$$cl(\phi)(q) = \sum_{S \models q} u \vee [\bigwedge_{p \in S} cl(\phi)(p)] + \neg \sum_{S' \models \neg q} u \vee [\bigwedge_{p \in S'} cl(\phi)(p)] \qquad (1)$$

For further details refer to [18].

5. LOW LEVEL MODULE

We formulate various rules for object equality and inequality based on our knowledge about scene structure and behavior of humans, vehicles and packages. These rules are applied to logical facts. Logical facts are generated by recognizing the occurrence of certain ground atomic events in video, such as an individual entering or exiting the field of view or dropping or picking up a package.

5.1 Object Detection and Local Tracking

Surveillance setups typically consist of cameras that are either fixed and observe the same scene at all times or cameras that can perform pan-tilt-zoom operations. Assuming static surveillance cameras gives us the advantage of being able to employ binary information obtained from background subtraction to detect objects of interest. We employ a background subtraction algorithm proposed in [10]. Tracking is performed by detecting foreground "blobs" in each frame and then matching them across consecutive frames using their color and spatial information. Across consecutive frames, these blobs either persist, merge with or split from other blobs, appear or disappear. We are concerned with tracking three kinds of objects - human, vehicles and packages (as opposed to tracking an arbitrary object like a hand). Due to a variety of reasons, background subtraction routinely introduces artifacts that can get tracked and erroneously labeled as objects of interest. Filtering out such noisy data is important for any tracking system; we do this by observing whether a blob persists across several frames or not.

While this form of temporal filtering culls out isolated blobs that might appear due to background subtraction errors, it does not remove regions comprised of pixels that deviate from the background model due to physical interactions between humans, vehicles, or packages and the scene. Examples are reflections and shadows that appear disconnected from the shadow-casting object. Filtering out of these artifacts can be assisted by the application of knowledge about the behavior of humans, vehicles and packages.

5.2 Fact Generation

Facts are generated when atomic events occur in the video including when objects appear, disappear, merge and split. These facts are annotated with named regions in the scene where they occur. These regions are manually labelled at setup. This helps us to generate relevant facts when an object interacts with that region in the image. From the point of view of tracking, we are primarily interested in regions in the image where objects appear and disappear from view. These regions typically correspond to scene boundaries, visible portals (doors to open or closed worlds) and static structures which could occlude tracked objects. Objects could also appear or disappear from the scene in areas other than the ones listed above. This usually happens when objects merge with or split from other tracked objects.

Objects can be occluded by static structures such as trees, pillars, boards etc. by moving behind them. We consider these occlusions to be closed worlds and expect that the objects will eventually emerge from them. An object disappearing into a closed world will cause the system to generate a fact $disappear(X, Closed\_World, T)$ where X is the object identifier, $Closed\_World$ is the identifier for the occlusion region and T is the time of the event. Similarly when an object appears from a closed world, a corresponding $appear$ fact will be generated. Tracked object interactions are also regarded as closed worlds and similar appear and disappear facts are generated when an object merges with or splits from another object.

Facts are also generated to record appearance matching scores between individuals. We employ two kinds of appearance matching algorithms. The first is a simple color histogram based algorithm. A histogram is constructed from the color values of the pixels in a segmented object and compared against color histograms of other objects using the Bhattacharyya distance. The second appearance matching algorithm is more sophisticated and is run using the feedback control mechanism described in the next subsection.

5.3 Control Feedback

When the high level module detects a deficit of information, it directs the low level module to gather information from certain space-time locales in the video. We maintain a queue of the previous 300 seconds of video along with the corresponding tracking data. This information also includes the segmentation information for each object detected.

Feedback in our system is of three types: (a) tracking back in time (b) searching for an object attached or contained within another object and (c) matching appearances of two individuals using spatial and color information. If the high level module requires tracking backward in time, it sends a "track_back" request to the low level module with the identifier of the object to be tracked and the time from which the tracking is to be done. We employ a mean-shift [3] based tracker which is initialized on the object to be tracked at the point it is detected and run backwards in time. A matching score is maintained and analyzed at each frame to ascertain whether or not the object being tracked is being tracked reliably. Tracking stops if either the queue of archived frames is exhausted or the matching score drops below threshold. In either case, the last tracked location is reported to the high level module.

If continuous tracking back in time is not possible e.g. if the target object itself is occluded as shown in Figure 1, then the high level module sends a "search" request with the frame number for searching and the identifier for the object of interest. Searching for an object attached or contained within the image of another object is performed using the standard cross correlation based hierarchical image matching algorithm. The search is carried out for 5 frames before and after the search frame requested by the high level module to avoid erroneous matches as far as possible.

The high level module can also request a higher complexity appearance matching between two individuals. It does this by sending a "match" request to the low level module which contains identifiers for both individuals it wants a matching score for. This appearance matching algorithm is the color path length based approach [22], which combines color and a geodesic path length measure within a person's body to construct a statistical appearance model. Models are compared using the Kullback distance. The path length for a pixel is the shortest distance measured along a path lying within the body, from the head to that pixel. See Figure 3, which displays the paths along which this distance is computed for a hand pixel. Note that although the Euclidean distance between the hand pixel and the head is different for the two poses, the geodesic distance stays nearly the same.

6. HIGH LEVEL MODULE

The primary task of the high level module is to reason about an object's identity when it appears from occlusions. It does so both in the bottom-up fashion, by taking input from the low level module and processing it, as well as in the top-down manner, actively seeking information where necessary. Reasoning is performed by applying predefined identity rules formulated in the MVDL framework to the set of facts generated by the low level module. When the truth value assigned to an identity decision is either unknown (u) or undecided ($*_{1\cdots n}$), the high level system determines if there exist any contextual cues that it can exploit. If it can, it provides control feedback to the low level module and directs it to collect historical information that will help it emerge from the unknown or undecided states of belief.

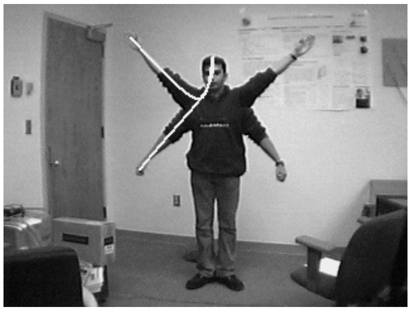

Figure 3: Figure showing geodesic path length for a hand pixel as measured from the head for two different poses.

6.1 Reasoning about identity

In [18], we employed four identifying cues or traits for reasoning about identities. These cues are based on the individual's possessions, closed world activity, knowledge and appearance. We continue to employ these identifying cues, although in a slightly different manner. In [18], we were primarily concerned with establishing identity across large visibility gaps, such as a person entering an office, and later re-appearing, or a person going around the corner into the open world and re-appearing. Due to the nature of the problem, we employed cues that would persist over that gap in time.

For example, the possession based rules state that identity can be verified on the basis of a person possessing something that only he can possess. So if it were known that a vehicle belonged to an individual and later another individual was observed entering that vehicle using a key that he possessed, it was concluded that the two individuals were equal. We were unable to conclude identity, however, based on less persistent objects like bags. For example, if an individual was observed to drop a bag in the scene and disappear from view and after a prolonged period of time, another individual was observed to appear in the scene and pick up the bag, it would not make sense to conclude that the two individuals were equal. The second individual could, for example, be committing a theft. However, in this work, since we are concerned with identity maintenance across occlusions, which are relatively short visibility gaps, we can exploit information provided by objects like bags to help establish identity. This requires that we make the assumption that bags do not change possession during the visibility gap. In addition to these four categories of rules, we also employ equality axioms of reflexivity, transitivity, and symmetry.

6.2 Identity Rules

It should be noted that any rule based on the cues listed above can almost never be definitive - most of them will be default rules. Also, different cues provide us with different amounts of information as they deal with varying degrees of uncertainty. Identity rules are formulated in the MVDL framework with 4 levels of priorities for defaults. Propositions can therefore assume values taken from the set

$B = \{u, dt_1, df_1, *_1, dt_2, df_2, *_2, dt_3, df_3, *_3, dt_4, df_4, *_4, t, f\}$. the bilattice for these set of truth values is shown in figure 4. The links shown in dotted lines indicate truth values for proposition derived from top-down active search of video. We assume that the definite rules (rules to which we assign truth values t and f) are always correct and therefore there can never be a contradiction between such rules. This assumption results in us ignoring the truth value $\perp$. Following we provide English descriptions of rules at each priority level.

Definite Rules: Definite rules are rules to which we assign truth value of either t or f, (i.e. we have the most confidence in the outcome of these rules). These rules capture knowledge that is always correct and that cannot be proven wrong (while most rules are default rules, definite rules act as stopping rules that terminate the revision of a proposition's belief state).

It is very hard to state that two individuals are definitely equal based on visual observation alone. Irrespective of how much information one includes in such rules, it is always possible to find ways to defeat them. Therefore, in our system we do not have a single rule that definitely infers equality. However, it is possible to state that two individuals are not equal. We do that when we observe them as two distinct individuals at the same instant of time. We also consider the equality axioms of reflexivity, transitivity and symmetry to be definite in nature.

Priority Level 4: Priority level 4 rules are those that compare only two individuals bound by either a closed world event or an identifying object. Occlusions are regarded as closed world events and therefore, an individual going out of view behind an occlusion is expected to reappear. We use this to formulate the following default rule: if we observe an individual enter an occlusion that we believe to be empty (no other individual is currently occluded there) and at a subsequent time, exiting it such that no other individual is observed to enter or exit that occlusion in the time between, then the two individuals are identical. If equality between two individuals is inferred using this rule, it will continue to hold as long as the system has no reason to believe that the occluded region was not empty during the period in question. If at anytime an individual that is not accounted for emerges from the occlusion, all identity assertions made until that point in time based on this rule are suspect and have to be retracted.

Other rules in this category are rules that state that if we observe an individual enter a occlusion and if, while we believe he is still behind, we observe another individual elsewhere in the scene, then these two individuals cannot be equal. Possession based rules also fall in this priority level. An individual can be said to be equal to another individual across an occlusion event if both of them are observed to carry a similar appearing object.

Priority Level 3: Priority level 3 rules are basically the possession based rules mentioned above. The only difference is that if identity is established based on actively searching for possession of a bag by way of top-down feedback (both searching and tracking back in time), we assign to it a truth value of priority level 3. The reason for placing feedback based possession rules one level below pure possession rules is because searching for similar looking objects attached to images of individuals is a less certain process. Our confidence in establishing identity based on what we *think* is a bag being carried by an individual is lower compared to knowing for certain that a bag was indeed carried. These rules are invoked only to resolve belief states of u, $*_1$ or $*_2$.

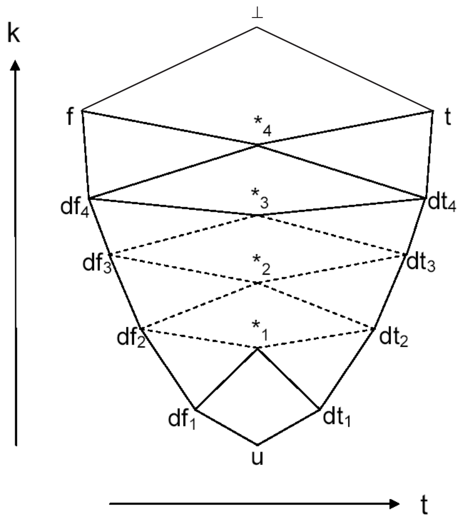

Figure 4: Bilattice employed in proposed system.

Priority Levels 2: Appearance matching based on color path length is assigned level 2. Appearance matching rules in general state that if two individuals appear similar, then they must be equal, while if the do not appear similar, then they must be not equal. Color path length based rules, while more accurate in matching appearance, are computationally expensive and are invoked by the high level module only when it is not possible to distinguish individuals based on the simple color histogram based appearance matching i.e. they are only used to resolve $*_1$ or u belief states.

Priority Level 1: Rules based on color histogram based appearance matching are assigned priority level 1 as these have the least information.

7. RESULTS

Our system has been implemented as a multi-threaded, C++ application. A Prolog reasoning engine has been embedded within this C++ application. Multivalued default reasoning is implemented using meta-predicates provided by Prolog. The application consists of two kinds of threads: the camera thread (the low level module), which take input from the camera and detects "atomic" events (like entering a door or picking up a bag) and a reasoning thread (high level module), responsible for the high level multivalued default reasoning. The camera thread first performs background subtraction and local tracking. It then detects "atomic" events and syntactically structures them as Prolog facts. The reasoning thread, when first created, starts the Prolog engine and initializes it by inserting into its knowledge base all the predefined rules from the default theory. The reasoning thread is subsequently evoked every few seconds. Every time it runs, it assimilates Prolog facts generated by the camera thread and inserts them into the Prolog engine's knowledge base. Also, for every human observed in the video, it reasons about their identity by applying all applicable equality rules. If it detects that any identity statement is in the "unknown" or any of the "undecided"

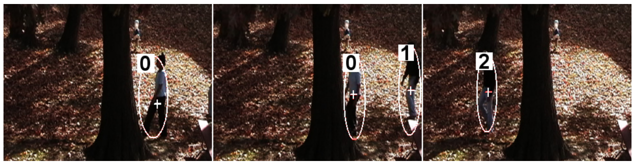

Figure 5: Sequence of images showing individual 0 and 1 disappearing from view and 2 subsequently appearing from behind an occlusion

states, it attempts to actively seek information to emerge from that state. All of the feedback controlled modules are run in a separate thread so they do not disturb the normal bottom-up functioning of the system.

7.1 Scenarios

We describe the scenarios used to test the system.

Scenario 1 (See Figure 5). *Two individuals 0 and 1 walk behind an occlusion. Individual 0 is wearing a blue shirt and black pants while individual 1 is wearing a black shirt and a blue pants. Individual 2 appears from the occlusion subsequently.*

In scenario 1, the overall color distribution between the two individuals 0 and 1 is similar and therefore, the level 1 rules, based on the color histogram based appearance matching, compute $\phi[\mathit{equal}(2,0)] = dt_1$ and $\phi[\mathit{equal}(2,1)] = dt_1$. However, the system is also able to prove that $\phi[\mathit{equal}(0,1)] = f$ and therefore by transitivity is forced to assign $\phi[\mathit{equal}(2,0)] = *_1$ and $\phi[\mathit{equal}(2,1)] = *_1$. Since the belief states of the identity statements is $*_1$, the high level module directs the system to use the level 2 appearance matching algorithm which employs color as well as spatial distribution of the pixels to match appearances. With this information, the system is now able to correctly conclude $\phi[\mathit{equal}(2,1)] = dt_2$ and $\phi[\mathit{equal}(2,0)] = \mathit{df}_2$.

Scenario 2 (See Figure 6). *Two individuals 0 and 1 approach each other and their views merge. Subsequently the individuals separate out and are now labelled 2 and 3 by the system. At this point, bag 4 is detected on the ground where 0 and 1 had merged. 3 exits the scene while 2 picks up bag 4 and exits the scene*

In this scenario too as in scenario 1, the system concludes $\phi[\mathit{equal}(3,0)] = *_1$, $\phi[\mathit{equal}(3,1)] = *_1$, $\phi[\mathit{equal}(2,0)] = *_1$ and $\phi[\mathit{equal}(2,1)] = *_1$. Application of level 2 rules does not help in this case, as both individuals are dressed alike and the system concludes $\phi[\mathit{equal}(3,0)] = *_2$,$\phi[\mathit{equal}(3,1)] = *_2$, $\phi[\mathit{equal}(2,0)] = *_2$ and $\phi[\mathit{equal}(2,1)] = *_2$. However when individual 2 picks up bag 4, the high level module can now potentially apply possession based level 3 set of rules. Therefore, it directs the low level module to track the bag backward in time from the point when it was first detected. The bag is correctly tracked back to individual 0 and the system concludes $\phi[\mathit{equal}(2,0)] = dt_3$.

Scenario 3 (See Figure 1). *Two similar looking individuals 0 and 1 disappear behind an occlusion and are completely lost sight of. Subsequently, individual 3 appears from the occlusion and drops bag 4 on the ground.*

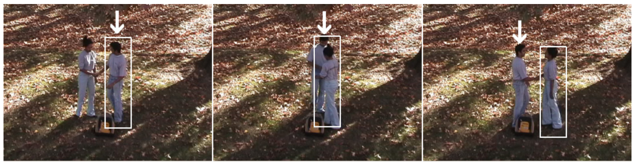

Figure 7: Typical tracking failure with conventional mean shift tracker. White arrow (manually inserted) shows correct track

As in the previous scenario, in this scenario too, since individuals 0 and 1 appear similar, both appearance matching cues are unable to assign to the identity statements any belief state greater than $*_2$. However, at the instant the bag 4 is detected, the high level module sends a search request to the low level module. The object to be searched for is the bag 4 and the objects to be searched within are images of 0 and 1. The system correctly identifies 1 as carrying the bag and revises $\phi[\mathit{equal}(3,1)]$ from $*_2$ to dt_3.

8. DISCUSSIONS AND SUMMARY

This paper described the use of context driven top-down and bottom-up image analysis for establishing identity of individuals across short visibility gaps. Use of the MVDL framework allows the system to use the degree of information of various cues to not only combine them in an information theoretic manner but to also detect situations where more information is needed and thus to drive the low level modules and actively seek information. Maintaining identity of individuals across occlusions is important for any surveillance application. Any system that employs conventional tracking algorithms will fail to handle situations such as those described in the scenarios in Section 7 (also see Figure 7). By giving the system the capacity to make identity decisions based on any available context, we give it the ability to handle some of these difficult cases. Understandably, not all occlusion events will be successfully handled by our system and most identity decisions will remain in “unknown” or “undecided” states. However, the fact that there exists a deficit of information will be explicitly known. This opens up possibilities for constructing more complex control feedback algorithms that can be used to extract more information from archival video that will help disambiguate.

9. REFERENCES

[1] C. BenAbdelkader, R. Cutler, and L. Davis. Motion-based recognition of people in eigengait space. In *Proc of Intl. Conf. on Auto Face and Gesture Recogtn*, page 267, 2002.

[2] G. J. Brostow and I. A. Essa. Motion based decompositing of video. *IEEE International Conference on Computer Vision*, 2001.

[3] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *IEEE Computer Vision and Pattern Recognition*, 2:142–149, 2000.

[4] M. L. Ginsberg. Multi-valued logics: a uniform approach to reasoning in artificial intelligence. *Computational Intelligence*, 4:265–316, 1988.

[5] I. Haritaoglu, D. Harwood, and L. Davis. W4: A real time system for detecting and tracking people. *IEEE*

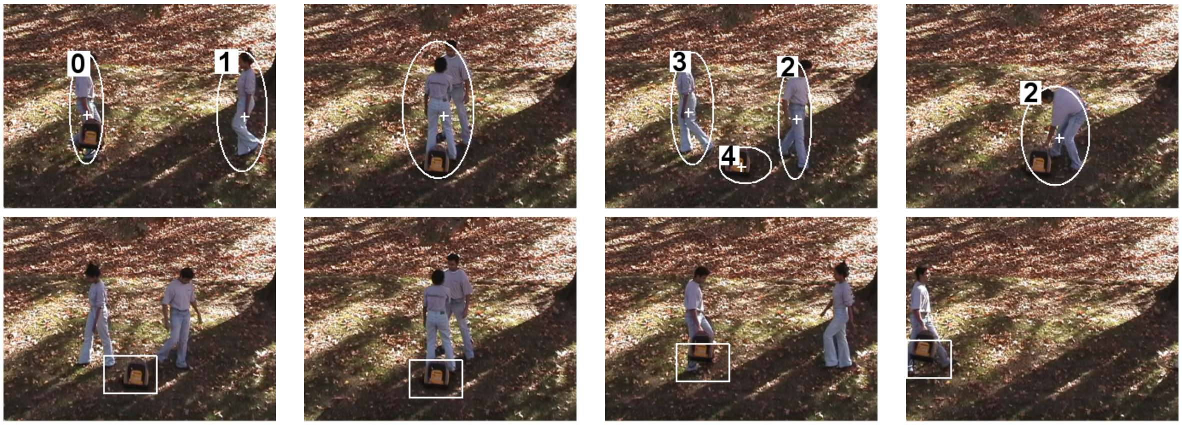

Figure 6: Figure described in scenario 2. Top row: Events detected bottom-up. Bottom row: Tracking back of bag

Computer Vision and Pattern Recognition, page 962, 1998.

[6] S. S. Intille and A. F. Bobick. Closed-world tracking. *IEEE International Conference on Computer Vision*, pages 672–678, 1995.

[7] M. Isard and J. MacCormick. Bramble: A bayesian multiple-blob tracker. *IEEE International Conference on Computer Vision*, pages 34–41, 2001.

[8] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25:1296–1311, 2003.

[9] N. Jojic and B. Frey. Learning flexible sprites in video layers. *IEEE Computer Vision and Pattern Recognition*, 2001.

[10] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis. Background modeling and subtraction by codebook construction. *IEEE International Conference on Image Processing*, 2004.

[11] D. Koller, J. Weber, and J. Malik. Robust multiple car tracking with occlusion reasoning. *European Conference on Computer Vision*, pages 189–196, 1994.

[12] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. Multi-camera multi-person tracking for easyliving. *Proc. 3rd IEEE Intl Workshop on Visual Surveillance*, 2000.

[13] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80:42–56, 2000.

[14] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern Recognition*, 36(9):1997–2006, 2003.

[15] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. *IEEE International Conference on Computer Vision*, 1995.

[16] R. Reiter. A logic for default reasoning. *Readings in nonmonotonic reasoning*, pages 68–93, 1987.

[17] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance models for occlusion handling. *In IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, 2001.

[18] V. Shet, D. Harwood, and L. Davis. Multivalued Default Logic for Identity Maintenance in Visual Surveillance. *European Conference on Computer Vision*, IV:119–132, 2006.

[19] H. Tao, H. S. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):75–89, 2002.

[20] G. Wei, V. Petrushin, and A. Gershman. Multiple-camera people localization in a cluttered environment. *The 5th International Workshop on Multimedia Data Mining*, 2004.

[21] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[22] K. Yoon, D. Harwood, and L. Davis. Appearance-based person recognition using color-path/length profile. Unpublished work.

[23] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Comput. Vis. Image Underst.*, 91(1-2):214–245, 2003.