Annotation Guidelines for Video Analysis and Content Extraction (VACE-II)

Version 6.2

Submitted to

Advanced Research and Development Activity

And

University of South Florida

January 25, 2006

Submitted by

VideoMining Corporation

(Formerly Advanced Interfaces Inc.)

Mr. Harish Raju
hraju@videomining.com
Mrs. Shubha Prasad
sprasad@videomining.com
Phone: (814) 867 8977
www.videomining.com

1. Introduction

This document explains the standards and practices followed by VideoMining Corporation (VM Corp, *formerly Advanced Interfaces Inc.*) during video annotation for the ARDA sponsored VACE-II research project. The tool being used to accomplish this task is the University of Maryland's Video Performance and Evaluation Resource (ViPER) application. For information on the use of ViPER refer to the documentation located at http://viper-toolkit.sourceforge.net/docs/gt/. While ViPER will be referred to often, one of the primary goals of this document is to provide detailed instructions capable of producing consistent and accurate ground truth across multiple annotators. Also this document is meant to serve as a reference for researchers making use of the annotated data. The procedures contained herein were developed over time through interaction between VACE-II technical monitors, the University of South Florida (USF), and the experiences of several annotators at Video Mining Corporation. Broad ranges of situations are covered with varying degrees of difficulty in order to keep annotations as precise as possible. In order to statistically annotate significant number of clips, annotation was decided to be done at I-Frame level .To be precise with the definition of I-Frame, it means annotation will be done for every 12th frame instead of doing it for each and every frame.

2. Annotation Tasks for VACE II

2.1 Task Overview

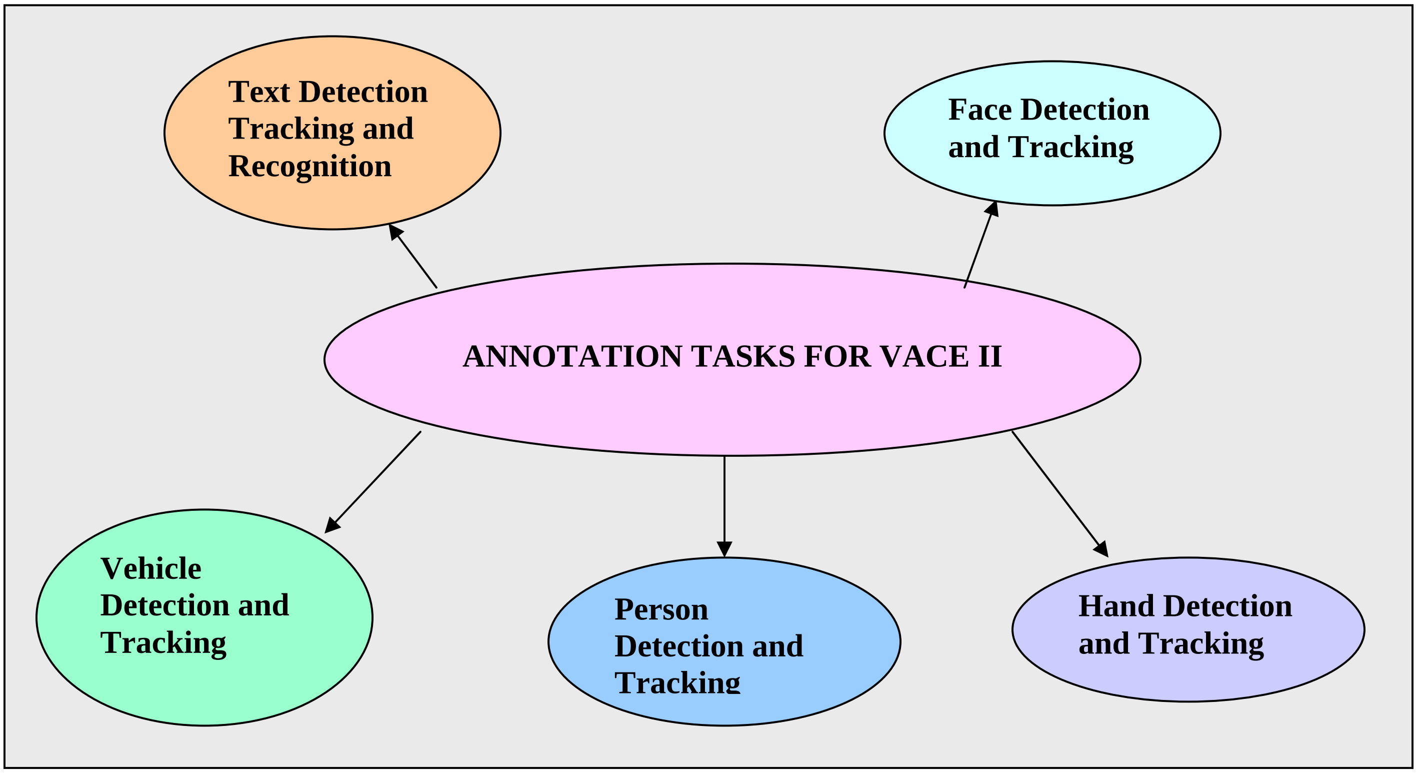

Text Detection Tracking and Recognition
Face Detection and Tracking
ANNOTATION TASKS FOR VACE II
Vehicle Detection and Tracking
Person Detection and Tracking
Hand Detection and Tracking

Fig. 2.1 - Overview of the Annotation Tasks for VACE II

2.1.1 Annotation Process Overview

As part of ensuring the best quality and consistency of annotation relevant to the VACE program a methodical approach has been adopted as detailed in Fig. 2.2. Based on the raw task definition obtained from the government panel and task focus groups, the VACE-CORE Eval-Panel provided a platform for open collaboration and refinement of the given tasks.

Based on the draft task guidelines, micro corpora of annotation were created to review the actual annotations. Upon successful acceptance of the micro corpora from the VACE members, the actual annotation tasks were carried out along with 10% of double annotation between different Image Data Analysts (IDA). In spite of the complexity and lack of robust analysis tools, a significant amount of effort was spent toward making the annotations consistent and accurate.

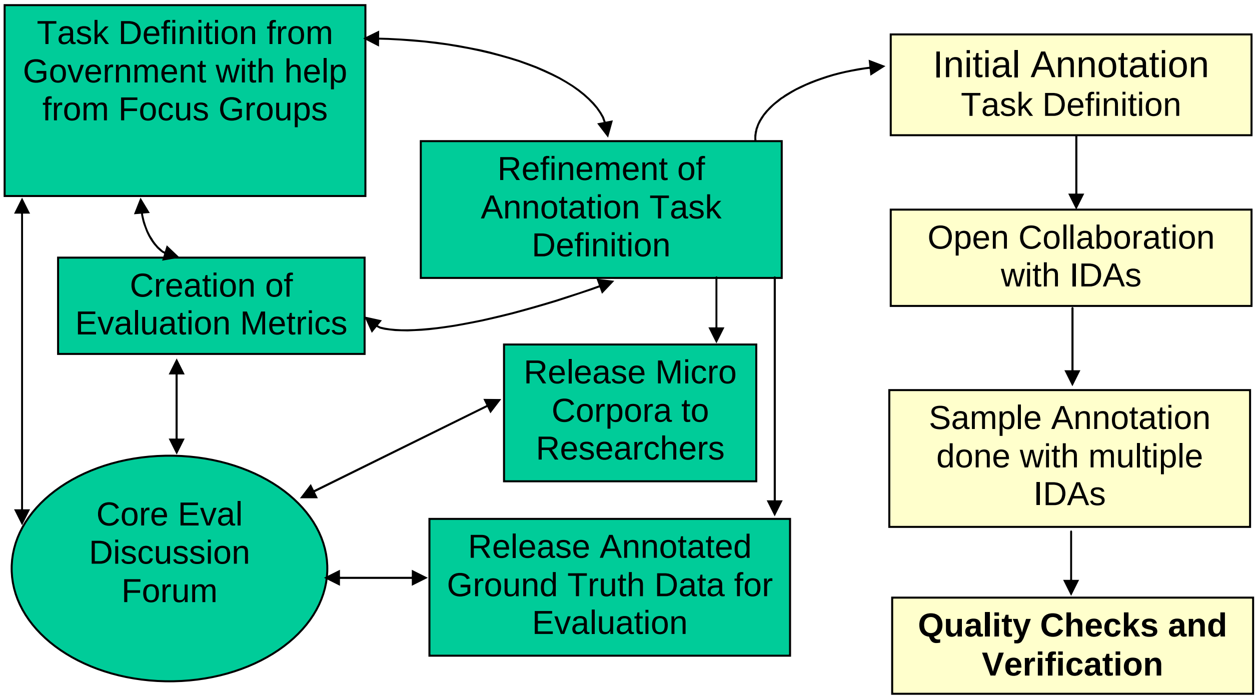

Task Definition from Government with help from Focus Groups
Refinement of Annotation Task Definition
Creation of Evaluation Metrics
Release Micro Corpora to Researchers
Core Eval Discussion Forum
Release Annotated Ground Truth Data for Evaluation
Initial Annotation Task Definition
Open Collaboration with IDAs
Sample Annotation done with multiple IDAs
Quality Checks and Verification

Fig. 2.2 - Overview of the Annotation Process

2.2 Annotation Task Matrix

Using ViPER, Annotation for both detection and tracking will be accomplished simultaneously. Hence, for the four domains namely `Meeting Room, Broadcast News, Unmanned Aerial Vehicles (UAV), and Surveillance`, we have five annotation tasks namely `Text, Face, Hand, Person, and Vehicle`. These video clips are selected based on the caveat that they will provide sufficient richness. However, we can rule out several task/domain pairs (vehicles in meetings, text, faces, and hands in UAV).

Domain Task	Meeting Room	Broadcast News	UAV	Surveillance
Text Detection, Tracking	⬛	🟩	⬛	⬛
Text Recognition	⬛	🟩	⬛	⬛
Arabic Text Detection, Tracking	⬛	🟩	⬛	⬛
Face Detection, Tracking	🟩	🟩	⬛	⬛
Hand Detection, Tracking	🟩	⬛	⬛	⬛
Person Detection, Tracking	🟩	⬛	🟩	🟩
Vehicle Detection, Tracking	⬛	⬛	🟩	🟩

Table 2.1 – Annotation Task versus Domain Matrix Status

Note:

a. All annotated domains include 10% doubly annotated data.
b. Each Task consists of Training and Testing Data. Typical data set included about 50 video clips.
c. Average Video Clip size is 2.5 minutes long.
d. RED = No Data.
e. GREEN = Task completed/In Progress.

2.3 Annotation Deliverables Schedule Estimate

The latest deliverables schedule is maintained on the NIST website. It will no longer be updated in the Annotation Guidelines document.

2.4 Individual Annotation Task Description

This section describes in detail the various annotation tasks along with their individual attributes and values. The following table lists the various data types used to represent the attributes.

Data Type	Description
BVALUE	Boolean Value. These variables hold TRUE / FALSE. The convention used is TRUE = 1 and FALSE = 0.
OBOX	Oriented Bounding Box. This is a 4 sided box with X, Y coordinates, W width, H height and O Orientation (0 to 360 Degrees).
DVALUE	Integer Value.
SVALUE	String Value.
FVALUE	Floating Value.
LVALUE	Sequence Value, can hold multiple items similar to a Drop Down List.
POINT	Point location. X and Y coordinates
ELLIPSE	Oriented Elliptical Bounding Box (X, Y, W, H, O)
BBOX	Non-Oriented Bounding Box. This is a 4 sided box with X, Y coordinates, W width and H height.

2.4.1 Frame Annotation

Attribute	Value	Description
[EVALUATE]	BVALUE	‘TRUE’ if frame is part of annotation; ‘FALSE’ if the frame is excluded from annotation.
[CROWD]	BVALUE	‘TRUE’ if the number of faces present >7; ‘FALSE’ if the number of faces present <=7.
[MULTIPLE VEHICLES]	BVALUE	‘TRUE’ if the number of vehicles >10; ‘FALSE’ if the number of vehicles <=10.
[MULTIPLE TEXT AREAS]	BVALUE	‘TRUE’ if number of lines of text >15; ‘FALSE’ if number if lines of text <=15.

2.4.2 Face Annotation

Attribute	Value	Description
[LOCATION]	OBOX	X, Y, H, W, O.
[VISIBLE]	BVALUE	‘TRUE’ if 1 eye, nose and part of mouth is seen; ‘FALSE’ otherwise.
[SYNTHETIC]	BVALUE	‘TRUE’ if its an artificial face; ‘FALSE’ otherwise.
[HEADGEAR]	BVALUE	‘TRUE’ if a person is wearing goggles /caps; ‘FALSE’ otherwise.
[AMBIGUITY FACTOR]	DVALUE	‘0’=Least confusing. When eyes, nose, mouth clearly seen; ‘1’=Partially confusing. When two out of three features are seen; ‘2’=Very confusing. When only one or none of the three features can be seen, including other confusing situations.

2.4.3 Text Annotation

Attribute	Value	Description
[LOCATION]	OBOX	X, Y, H, W, O.
[TYPE]	DVALUE	0 for Graphic Text; 1 for Scene Text.
[READABILITY]	DVALUE	0 if text is completely unreadable; 1 if text is partially readable; 2 if text is clearly readable.
[CONTENT]	SVALUE	Transcribed when [READABILITY] is set to 2, or [OCCLUSION] is 'FALSE' or when [SPECIAL TEXT] is 'FALSE'. Null when [READABILITY] is set to 0 or 1, or [OCCLUSION] is 'TRUE' or when [SPECIAL TEXT] is 'TRUE'. Transcribed with 'FR' when annotating Foreign text or language.
[OCCLUSION]	BVALUE	'TRUE' if text is occluded; 'FALSE' otherwise.
[LOGO]	BVALUE	'TRUE' if text is a company or business symbol / logo; 'FALSE' otherwise.
[SPECIAL TEXT]	BVALUE	'TRUE' if its non standard text and does not satisfy Graphic or Scene Text definition; 'FALSE' otherwise.
[DYNAMIC TEXT]	BVALUE	This flag is valid only when [SPECIAL TEXT] is TRUE. 'TRUE' if text changes dynamically instead of periodic changes like scrolling / ticking. 'FALSE' if text is scrolling or ticker based.

2.4.4 Hand Annotation

Attribute	Value	Description
[LOCATION]	POINT	X, Y point indicating the center for the palm area of the hand.
[VISIBLE]	BVALUE	'TRUE' if hand is present in the scene.
[SYNTHETIC]	BVALUE	'TRUE' if its an artificial hand like a cartoon hand or statue hand; 'FALSE' otherwise.
[HANDGEAR]	BVALUE	'TRUE' if a person is wearing gloves or other hand accessories covering the palm and fingers; 'FALSE' otherwise.
[LARGE HAND]	OBOX	If the hand occupies more than 25% of the frame then the hand is annotated using the OBOX.
[OCCLUSION]	BVALUE	'TRUE' if any part of hand is occluded by something other than itself; 'FALSE' otherwise.

2.4.5 Moving Vehicle Annotation

2.4.5.1 UAV Test Data Set

Attribute	Value	Description
[LOCATION]	OBOX	X, Y, H, W, O.
[VISIBLE]	BVALUE	‘TRUE’ if vehicle is present in the scene.
[OCCLUSION]	BVALUE	‘TRUE’, if vehicle is hidden behind other vehicles or objects; ‘FALSE’ otherwise.
[CATEGORY]	DVALUE	0 if car or sedan; 1 if trailer truck; 2 for any other type of vehicle.
[GROUP OF VEHICLES]	BVALUE	If more than 7 vehicles are parked together then the vehicles are grouped together using an OBOX.
[MOBILITY]	LVALUE	‘Mobile’ if the vehicle is moving; ‘Stationary’ if the vehicle is parked or standing still.

2.4.5.2 UAV Train Data Set

Attribute	Value	Description
[LOCATION]	OBOX	X, Y, H, W, O.
[VISIBLE]	BVALUE	‘TRUE’ if vehicle is present in the scene.
[OCCLUSION]	BVALUE	‘TRUE’, if vehicle is hidden behind other vehicles or objects; ‘FALSE’ otherwise.
[CATEGORY]	DVALUE	0 if car or sedan; 1 if trailer truck; 2 for any other type of vehicle.
[GROUP OF VEHICLES]	BVALUE	If more than 7 vehicles are parked together then the vehicles are grouped together using an OBOX.

2.4.5.3 Surveillance Domain

Attribute	Value	Description
[LOCATION]	BBOX	X, Y, H, W.
[PRESENT]	BVALUE	‘TRUE’ if Vehicle is visible in the scene.
[OCCLUSION]	BVALUE	‘TRUE’, if vehicle is hidden behind another object in the frame eg cars, trees etc ‘FALSE” otherwise.
[CATEGORY]	DVALUE	0 if car or sedan, 1 if trailer truck, 2 for any other type of vehicle.
[AMBIGUOUS]	BVALUE	If a particular region/car is difficult to annotate, Ambiguity is set to ‘TRUE’, ‘FALSE’ otherwise.
[MOBILITY]	BVALUE	‘Mobile’ if the vehicle is moving, ‘Stationary’ if the vehicle is parked or standing still.

2.4.6 Person Annotation

2.4.6.1 Meeting Room Domain

Attribute	Value	Description
[HEAD LOCATION]	ELLIPSE	X, Y, H, W, O (Oriented Elliptical Box as applicable in ViPER).
[BODY LOCATION]	OBOX	X, Y, H, W, O (Rectangular Oriented Bounding Box).
[VISIBLE]	BVALUE	'TRUE' if person is present in the scene.
[SYNTHETIC]	BVALUE	'TRUE' if artificial person for e.g. cartoon or statue; 'FALSE' otherwise.
[OCCLUSION]	BVALUE	'TRUE' when person is hidden behind objects or by other people; 'FALSE' otherwise.
[HEADGEAR]	BVALUE	'TRUE' if a person is wearing a hat, helmet, or clothing which obscures the head. Spectacles and Glasses will not be considered as [HEADGEAR].
[AMBIGUOUS]	BVALUE	'TRUE' if the person cannot be bound due to camera view or extreme [OCCLUSION]. e.g. when a person is cut off from camera view; 'FALSE' otherwise.

2.4.6.2 UAV Domain

Attribute	Value	Description
[LOCATION]	OBOX	X, Y, H, W, O (Oriented Bounding Box).
[VISIBLE]	BVALUE	'TRUE' if person is present in the scene.
[SYNTHETIC]	BVALUE	'TRUE' if artificial person, for e.g. cartoon or statues; 'FALSE' otherwise.
[OCCLUSION]	BVALUE	'TRUE' when person is hidden behind other objects or people; 'FALSE' otherwise.
[HEADGEAR]	BVALUE	'TRUE' if a person is wearing a hat, helmet, or clothing that obscures the head. Spectacles and Glasses will not be considered as [HEADGEAR].
[AMBIGUOUS]	BVALUE	'TRUE' if the person cannot be bound due to camera view or extreme [OCCLUSION]. e.g. when a person is cut off from camera view; 'FALSE' otherwise.
[MOBILITY]	LVALUE	'Mobile' when the person is moving/walking; 'Stationary' if the person is standing still and not moving.

2.4.6.3 Surveillance Domain

Attribute	Value	Description
[LOCATION]	BBOX	X, Y, H, W
[PRESENT]	BVALUE	'TRUE' if vehicle is visible in the scene.
[OCCLUSION]	BVALUE	'TRUE', if vehicle is hidden behind another object in the frame eg cars, trees etc. 'FALSE' otherwise.
[SYNTHETIC]	BVALUE	'TRUE' if person present is not real eg statues, cartons etc. 'FALSE' Otherwise
[AMBIGUOUS]	BVALUE	If a particular region/car is difficult to annotate, Ambiguity is set to 'TRUE', 'FALSE' otherwise.
[MOBILITY]	BVALUE	'Mobile' when the person is moving/walking; 'Stationary' if the person is standing still and not moving.

2.4.7 *FILE* Descriptor

Each time a metadata file is created in ViPER a FILE descriptor is generated. Its purpose is to collect data concerning the media used in the annotation. No user interaction is required as the appropriate values are detected and entered automatically.

Attribute	Type	Description
[SOURCE TYPE]	LVALUE	'Sequence' or 'Frames' (automatic).
[NUMFRAMES]	DVALUE	Integer value (automatic).
[FRAMERATE]	FVALUE	Floating point value (automatic).
[H-FRAME-SIZE]	DVALUE	Integer value (automatic).
[V-FRAME-SIZE]	DVALUE	Integer value (automatic).

2.4.8 Default Annotation Attributes

Each descriptor in ViPER has several attributes associated by default. These are universal to any object and cannot be removed. The first attribute is labeled in ViPER simply as "P" for [PROPAGATE]. This attribute should only be checked when used to copy values through a range of frames, otherwise it is left blank. The [VALID] attribute is used to denote when an object is active in the video**.** [VALID] is checked from the frame that it first appears through the last frame before the object is out of view. The last default attribute is the object [ID]. A unique value is generated automatically and placed in the [ID] each time an instance is created. Assignment is handled completely by ViPER; therefore this is not an editable attribute.

Attribute	Value	Description
[PROPAGATE]	CHECKBOX	"Checked" when copying values to range of frames.
[VALID]	CHECKBOX	"Checked" when object appears until exits video.
[ID]	DVALUE	Integer value (automatic).

3. Annotation Tasks Description

This section discusses the aspects of frame annotation including special cases and sample annotation frames.

3.1 Frame Annotation Attributes

3.1.1 [EVALUATE]

The FRAME descriptor has four attributes: [EVALUATE], [CROWD], [MULTIPLE VEHICLES], and [MULTIPLE TEXT AREAS]. For each video, there will be segments that are too difficult to annotate. An example is a clip in which the camera is moving quickly, making the faces blurred, or when there is significant interference in the video. Such video segments will be flagged as [EVALUATE] = FALSE, while clear video will be [EVALUATE] = TRUE. In addition, transitional scenes are particularly difficult and time consuming to annotate. For this reason, scenes that contain one scene fading into another could be excluded from annotation and flagged [EVALUATE] = FALSE. Fig. 3.1 shows an example of this annotation rule.

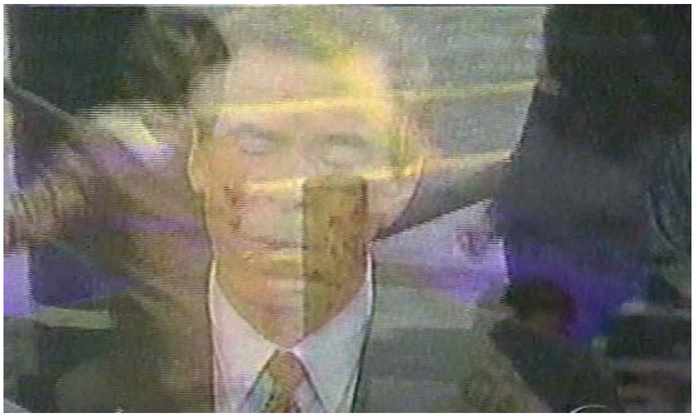


Fig. 3.1 – A transition where [EVALUATE]=*FALSE*

3.1.2 [CROWD]

There are video segments in which too many objects are in the scene to mark and track. One example would be people in a stadium or a group of soldiers. In such cases when the number of faces exceeds seven, the annotator will tag the frame by setting [CROWD] = TRUE as seen in Fig. 3.2. The annotator will not be required to annotate the frame when the value of [CROWD] is set to TRUE. Otherwise, while [CROWD] = FALSE, the annotator will mark and track each face in the image. Fig. 3.2 shows a picture when the [CROWD] attribute is set to 'TRUE'.

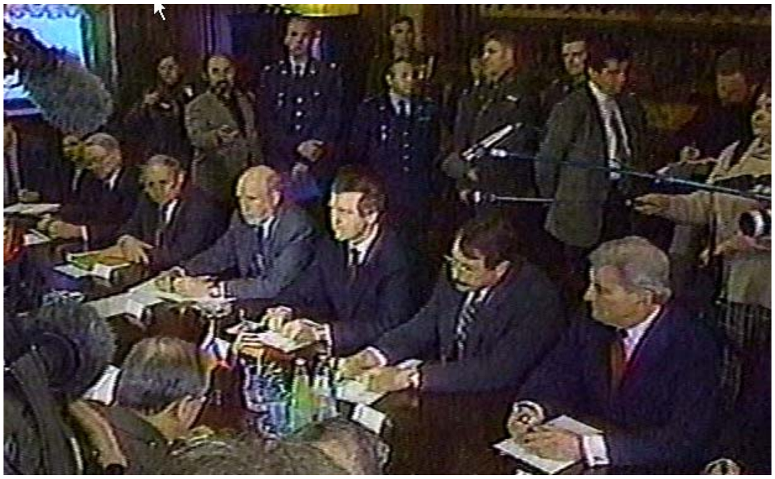


Fig. 3.2 - Example of a video clip where the number of faces exceeds 7, hence [CROWD] = TRUE

3.1.3 [MULTIPLE VEHICLES]

When there are many valid vehicles (15 or more) present in a scene, [MULTIPLE VEHICLES] is set to TRUE. As with the [CROWD] attribute, the annotator is not required to mark or annotate vehicles when [MULTIPLE VEHICLES] is set to TRUE.

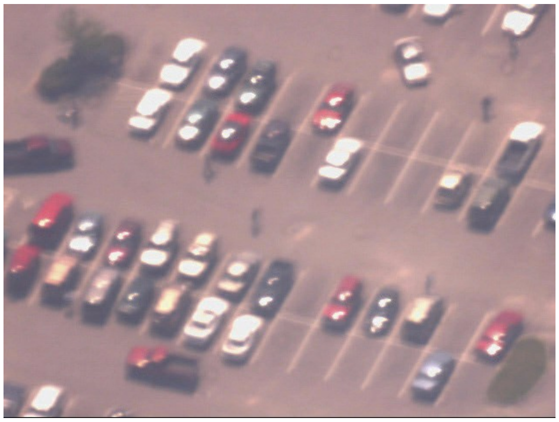


Fig. 3.3 - Example of a video clip where the number of vehicles exceeds 15, hence [MULTIPLE VEHICLES] = TRUE

3.1.4 [MULTIPLE TEXT AREAS]

Similar to the previous two attributes, [MULTIPLE TEXT AREAS] defines frames where the numbers of annotated text areas are greater than fifteen. A text area is considered to be a rectangular bounded box containing text of similar readability level, size, and font. Further rules for defining a bounding box are listed in section 3.3.

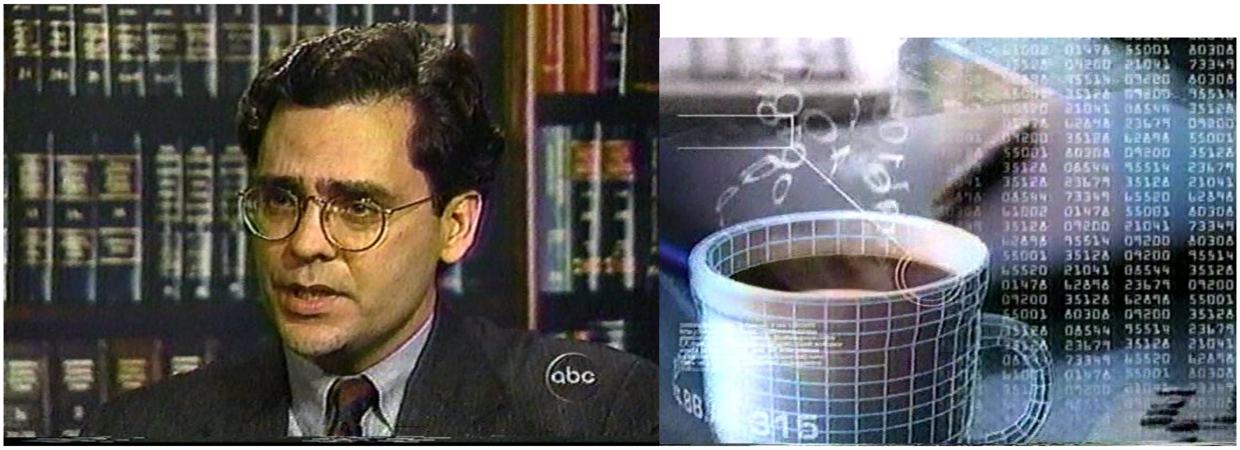


Fig. 3.4 – Examples of [MULTIPLE TEXT AREAS] containing too much text for annotation. Hence the [MULTIPLE TEXT AREAS] is set to ‘TRUE’

3.2 Face Annotation Task

Every new face is marked with a box when it appears. Annotation of a face does not begin until at least one eye, the nose, and part of the mouth is present in the scene. Moving and scaling the box as the face moves in the succeeding frames tracks the face. This process is done until the person disappears from the frame. There may be times when the face is seen and times when it is not seen. The [VISIBLE] attribute is set to TRUE when the face is seen and FALSE otherwise. The face will be continued to be tracked with approximate bounding region and [VISIBLE] will be set to FALSE, when an object such as a tree temporarily blocks a face.

If a face moves out of frame and returns, it is treated as a new instance rather than the same one. In addition, if the camera perspective changes suddenly, as happens from switching angles or magnification, then the face is treated as a new instance.

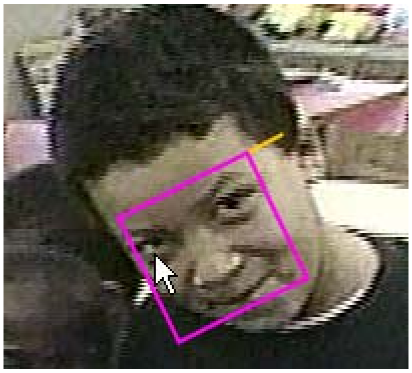


Fig. 3.5 – Annotated skewed face

3.2.1 Face Task [VISIBLE] Attribute

A face has [VISIBLE] attribute set to TRUE if at least one eye, the nose and part of the mouth is seen. Annotations can be done on profile views if they satisfy these constraints.

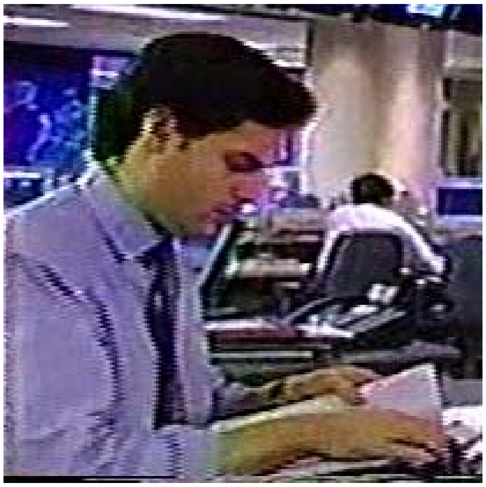
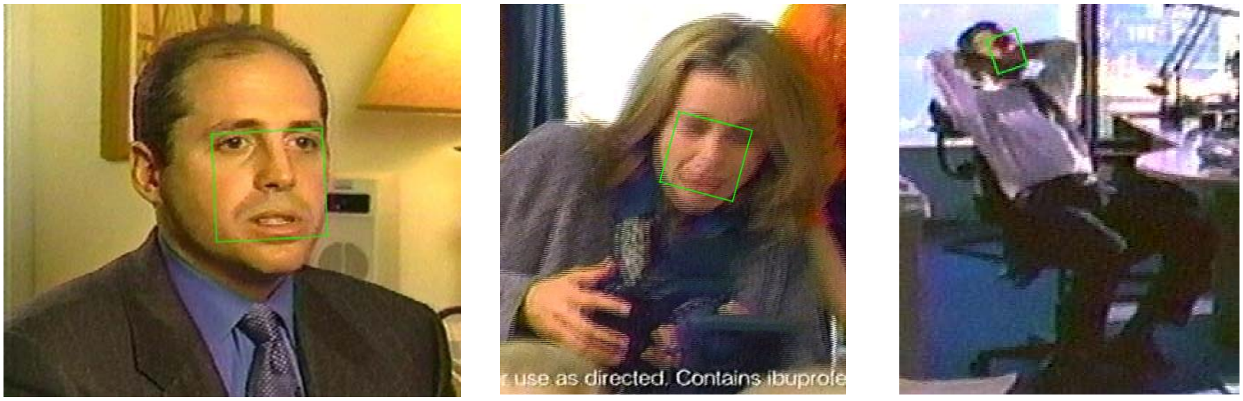


Fig. 3.6 – Profile face, considered [VISIBLE] = TRUE

In situations when a person is wearing a cap, goggles, etc, annotation is still required for the face as we have an approximation as to where the eye is. These cases are annotated the same as regular faces but with an additional attribute [HEADGEAR] set to TRUE.
[HEADGEAR] will be TRUE for any case in which a person is wearing glasses or any head cap or garment that covers the eyes. However, this does not include when a mask covers a face. For instance, a ski mask hiding nearly all facial features is considered as [VISIBLE] =FALSE.

Often faces are seen in circumstances that make judging visibility difficult. These include instances of faces in the distance, poor quality video, objects occluding the face, etc. For this reason, the attribute [AMBIGUITY FACTOR] is used to categorize the level of certainty in which a face is annotated. A face with eyes, nose, and mouth clearly seen is given the attribute [AMBIGUITY FACTOR] a value of 0. A face where the features are seen but not clear or when only two of the three features (eyes, nose, and mouth) are seen the attribute [AMBIGUITY FACTOR] is given a value of 1. Lastly a situation in which video quality is very poor, hands occlude the face, one out three features or less are [VISIBLE], or the situation is otherwise confusing, then the attribute [AMBIGUITY FACTOR] is given a value of 2. Examples for all three cases can be seen in Fig. 3.6.1.



[AMBIGUITY FACTOR] = 0
All facial features are visible

[AMBIGUITY FACTOR] = 1
Motion blur distorts face

[AMBIGUITY FACTOR] = 2
Facial features barely discernable

Fig. 3.6.1 – Examples of [AMBIGUITY FACTOR] levels

3.2.2 Guidelines for Face Region Bounds

A face is bounded with an oriented box. Features of the face will be used as guides for marking the limits of the box edges. If the features are obstructed (e.g. sunglasses, hat, etc.) then the marks are approximated. The top of the box is drawn such that it is above the eyes and below the eyebrows. The bottom is drawn to be above the front of the chin and below the lower lip. The sides of the box should be tight such that the edges enclose the outermost sides of the eyes.

Let D be the distance from the point of the nose to the center of the mouth. When marking a face, the space between the facial features (eyes, nose, and mouth) and the sides of the selection box should not exceed D / 2. Note that, for small faces, an approximate bounding box is drawn.

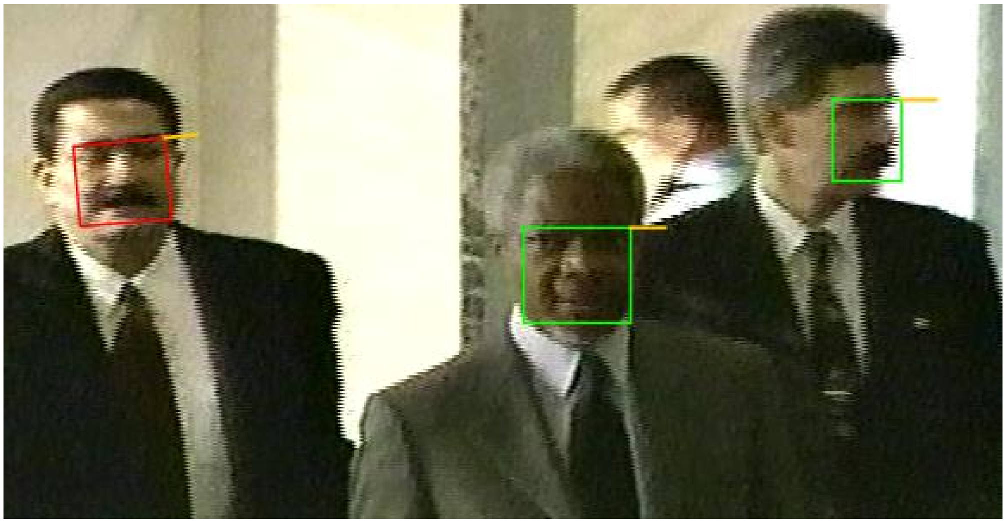


Fig. 3.7 –Examples of Face annotation

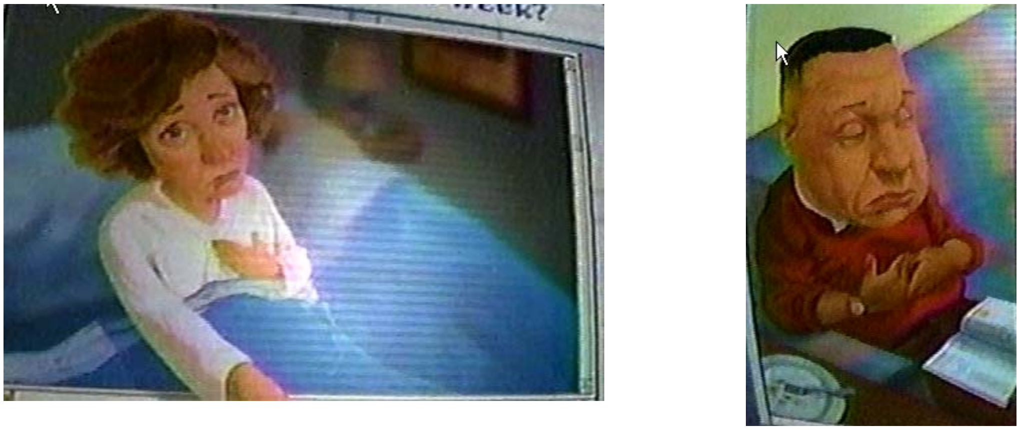


Fig. 3.8 – Examples of `[SYNTHETIC]` Faces

Some videos will contain artificial faces such as statues, paintings, or cartoons. When this occurs, the face is to be annotated with the attribute `[SYNTHETIC]` = `TRUE`. All other faces will have `[SYNTHETIC]` = `FALSE`.

3.3 Text Annotation Task

Every new text area is marked with a box when it appears. Moving and scaling the selection box tracks the text as it moves in succeeding frames. This process is done until the text disappears from the frame. There are two types of text:
1. Scene Text
2. Graphic Text

3.3.1 Scene Text and Graphic Text

Scene text is anything in the background of what is actually being filmed.
Examples are:
- Words on the bottle of water that someone is drinking,
- Words on a poster that is within the picture,
- Words on a sign that is part of a picture,
- Words on a candy bar wrapper that is part of the picture.

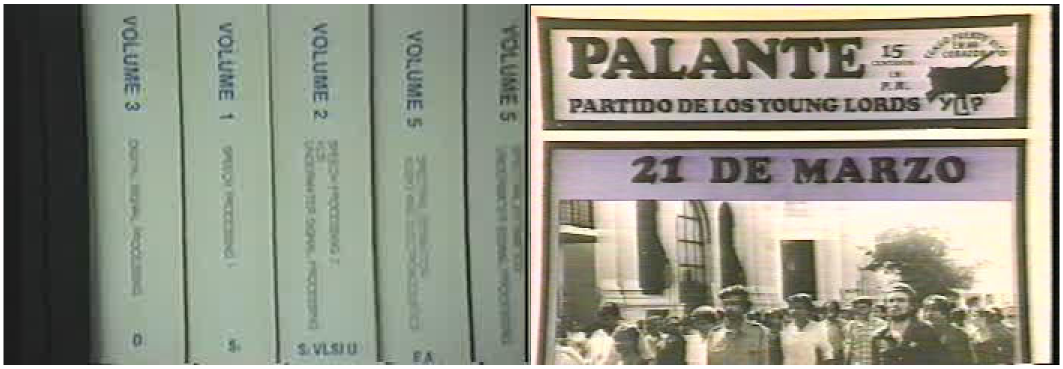


Fig. 3.9 – Examples of Scene Text

Graphic text is anything overlaid into the picture.
Examples are:
- Television channel numbers,
- Television captions,
- Sports score updates.

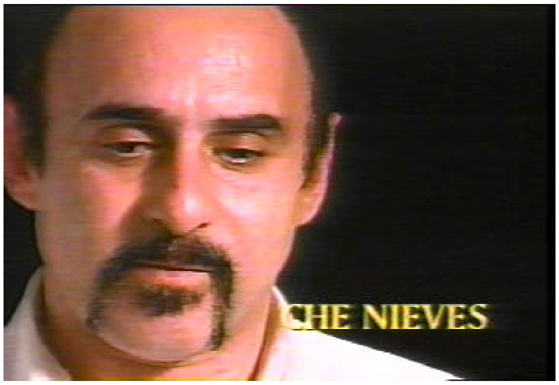


Fig. 3.10 – Example of Graphic Text

3.3.2 Guidelines for Text Region Bounds

Text is particularly difficult to be uniformly bound. For this reason, text bounding will be performed using a set of rules. The first rule is that all text within a selected block must contain the same [READABILITY] level and [TYPE]. The second rule is that blocks of text must contain the same size and font. Two allowances are given to this rule. A different font or size may be included in the case of a unique single character and also the font color may vary among text in a group. The third rule is that the bounding box should be tight to the extent that there is no space between the box and text. The maximum distance from the box to the edge of bounded text may not exceed half the height of the characters when [READABILITY] = 2 (clearly readable). When [READABILITY] = 0 or 1 the box should be kept tight but does not require separate blocks for partial lines in a paragraph. For an example refer to Fig. 3.11. The final rule is that text boxes may not overlap other text boxes unless the characters themselves are specifically transposed atop one another.

To summarize for bounding a text area:

- Text attributes must be consistent throughout the block

- Text font and size must be consistent throughout the block, except when only one letter is different from the group (color may vary)

- The bounding box should leave no space between the box and text edges.

- Distance from bounding boxes to edges of enclosed text may not exceed ½ the average character height when [READABILITY] = 2. When [READABILITY] = 0 or 1 the space from the partial line of a paragraph is included in a single block. Text boxes may not overlap unless the text itself is layered

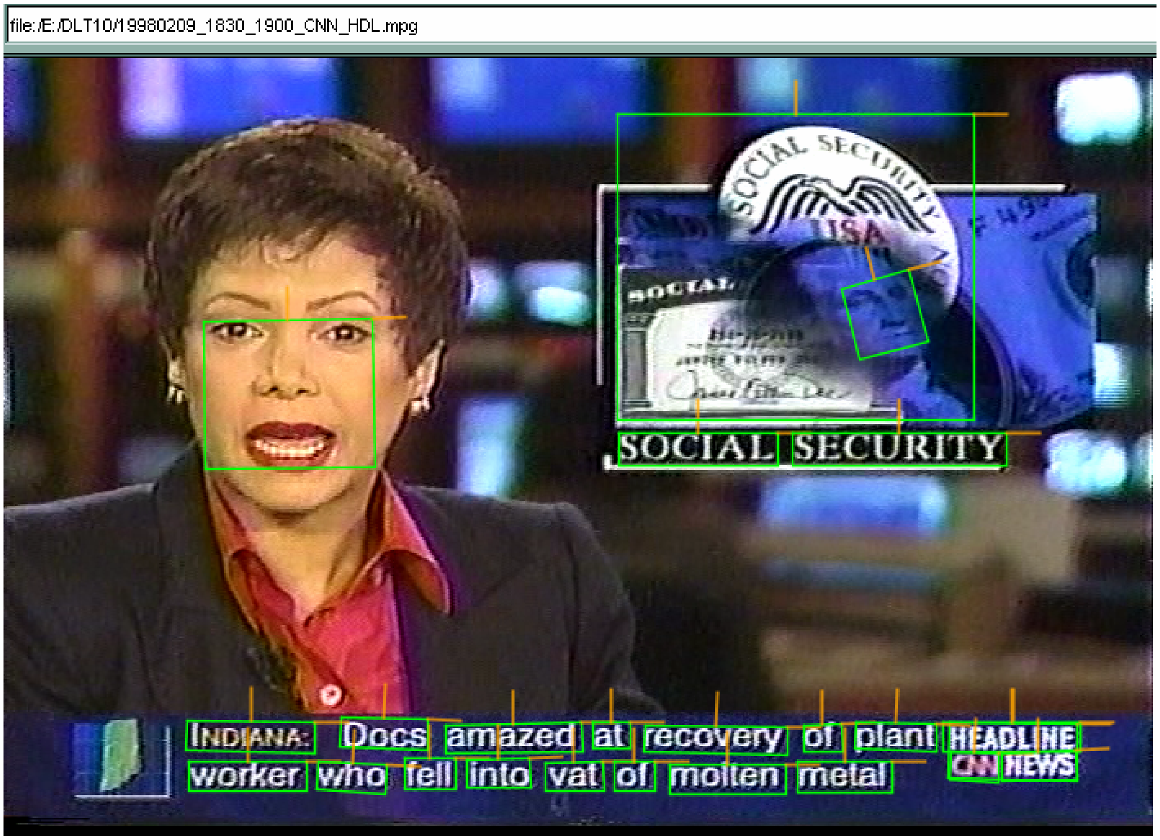

file:/E:/DLT10/19980209_1830_1900_CNN_HDL.mpg
SOCIAL SECURITY
SOCIAL SECURITY
USA
SOCIAL
INDIANA: Docs amazed at recovery of plant
worker who fell into vat of molten metal
HEADLINE
CNN NEWS

Fig. 3.11 – Examples of Text Annotation

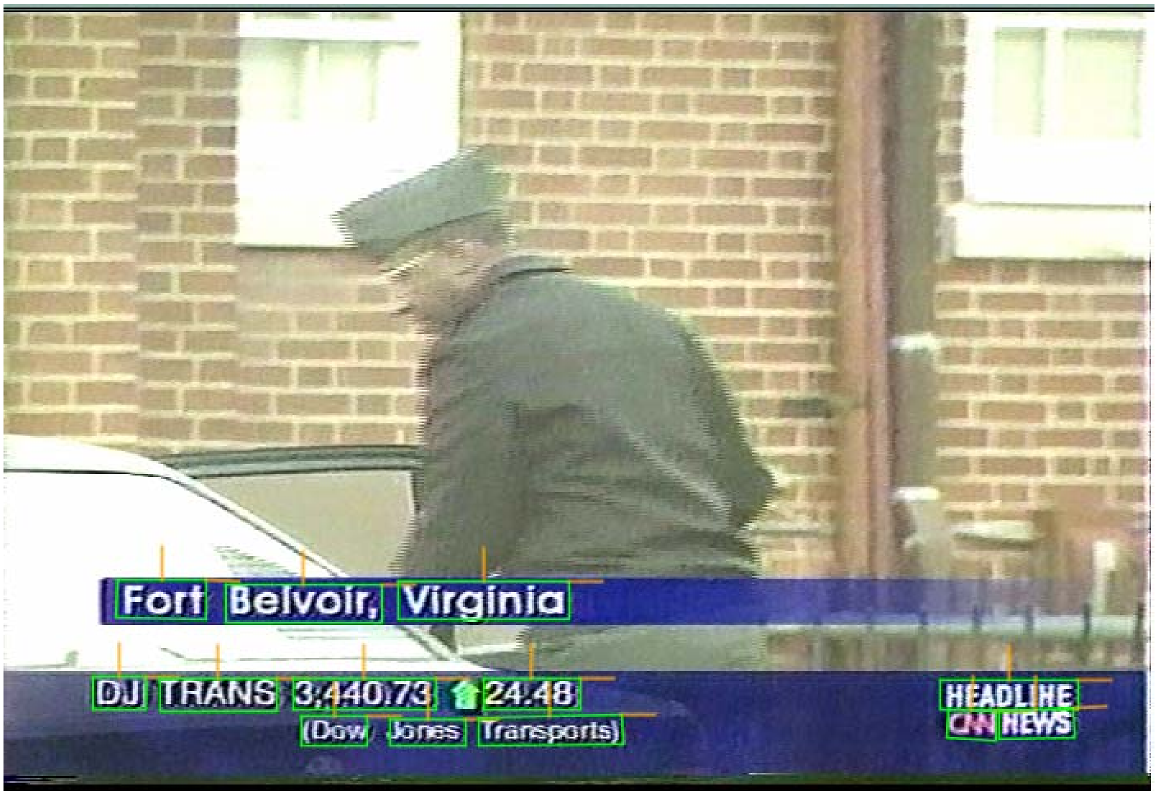


Fig. 3.12 – Example of text annotation with CNN marked as a [LOGO]

When the text is cut off by the bounds of the frame or by another object, the text is said to be Occluded. When the text is occluded we set the [OCCLUSION] = TRUE, else it is set to FALSE.

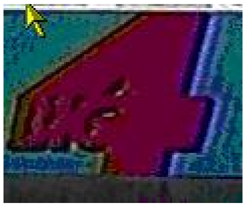


Fig. 3.13 – Text within text. This is annotated as two separate blocks.

3.3.3 Guidelines for Text [READABILITY] Attribute

[READABILITY] attribute consists of three levels. Completely unreadable text is signified by [READABILITY] = 0 and is defined as text in which no character is not identifiable. Partially readable text is given [READABILITY] = 1 and contains characters that are both identifiable and non-identifiable. Clearly readable text is assigned [READABILITY] = 2 and is used for text in which all letters are identifiable.

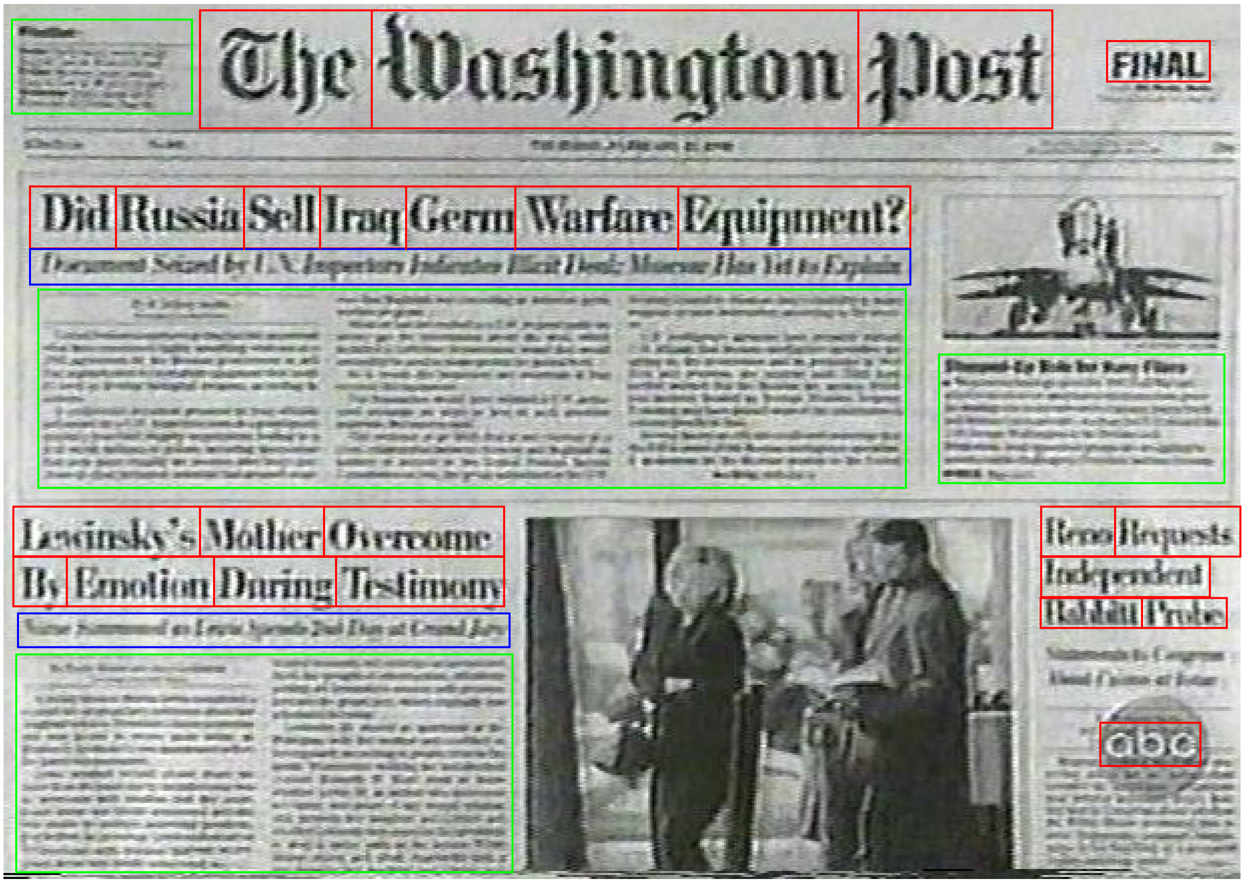

The Washington Post
FINAL
Did Russia Sell Iraq Germ Warfare Equipment?
Lewinsky's Mother Overcome
By Emotion During Testimony
Reno Requests
Independent
Rabbit Probe
abc

Fig. 3.14 - Example of many text [READABILITY] levels, note the "abc" graphic covering scene text

Red Lines indicate [READABILITY] = 2 and will be transcribed
Blue Lines indicate [READABILITY] = 1 and will not be transcribed
Green Lines indicate [READABILITY] = 0 and will not be transcribed

3.3.4 Text Recognition

Text Recognition task inherits most of the high level attributes of the Text Detection task with the following modifications / additions:

1. If the [READABILITY] = 2, the text annotation will be done at the word level. For e.g. in Fig. 3.14, for sections that have [READABILITY] = 2, individual lines will be picked for word level annotation.

2. VIPER will allow for conversion of line level text transcription into Word Level Objects. This is illustrated in Fig. 3.14. The original line level annotations with [READABILITY] = 2 are shown with their respective Word level objects.

3. All Word level objects will have the same attributes as the current Text Objects. Refer to Table 2.4.3 for complete description.

4. [OCCLUSION] and Non Linear Text will be handled differently than line level text task. As you can see in Fig. 3.15a, occluded words will not transcribed but clearly

[VISIBLE] words will be transcribed. In Fig. 3.15b, non linear words are handled by individually annotating each word using oriented bounding boxes.

5. The [CONTENT] attribute is transcribed when the [READABILITY] attribute is set to 2, or the [OCCLUSION] is 'FALSE', or the [SPECIAL TEXT] is 'FALSE'.

6. The [CONTENT] attribute is Null when [READABILITY] attribute is set to 0 or 1, or the [OCCLUSION] is 'TRUE', or the [SPECIAL TEXT] is 'TRUE'.

Note: In Foreign Text Annotation the [CONTENT] Attribute will be transcribed as "FR'. Any other language such as English text in the Arabic Text Domain will have the data transcribed for the [CONTENT] attribute.

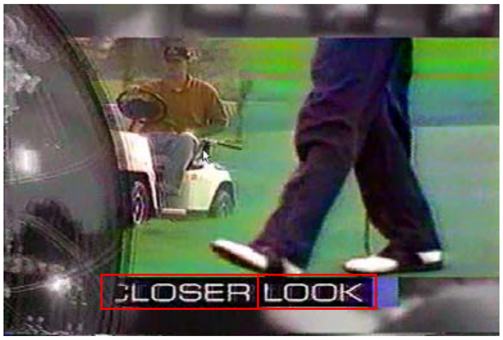


Fig. 3.15a – Occluded text due to overlaying graphic will not be transcribed. Here, the word "CLOSER" will not be transcribed, but "LOOK" will be transcribed

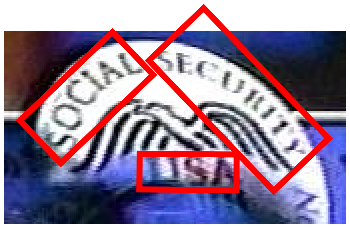


Fig. 3.15b – Non Linear Text regions will be transcribed at word level using Oriented Bounding boxes. Here, SOCIAL, SECURITY and USA are transcribed as shown.

3.3.5 Special Cases

Text may have properties or behaviors too challenging for the scope of this annotation task. Text exhibiting these qualities is flagged as a [SPECIAL TEXT]. A [SPECIAL TEXT] should be used when text is continuously scrolling or unreadable. Examples include continuously scrolling text as seen in Fig 3.16, sports scores ,etc. The annotation guidelines cannot cover every possible text scenario. Therefore, a [SPECIAL TEXT] may also be used as a "catch-all" for undefined text cases that the annotator is unsure how to properly annotate. In addition when [SPECIAL TEXT] = TRUE, [DYNAMIC TEXT] attribute is set to FALSE to indicate that the text is scrolling or ticker based. Else, [DYNAMIC TEXT] is set to TRUE to indicate irregular changes in text content because of transition or graphical animation.

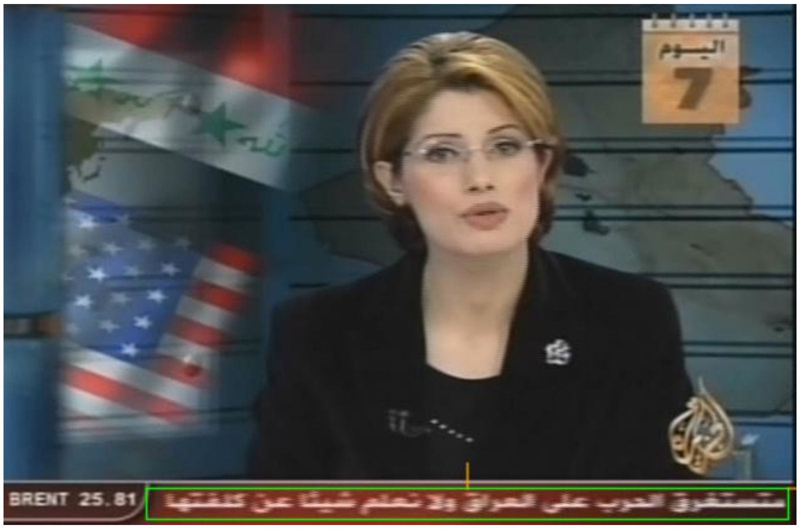


Fig. 3.16 – Continuously scrolling text is to be flagged as
`[SPECIAL TEXT]` = `TRUE ; [DYNAMIC TEXT] = FALSE`

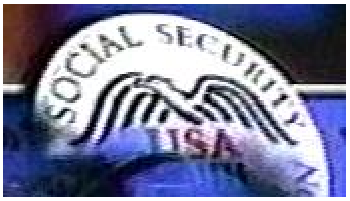


Fig. 3.17 – Partially readable text will be marked but not transcribed
`[SPECIAL TEXT]` = `TRUE ; [DYNAMIC TEXT] = TRUE`

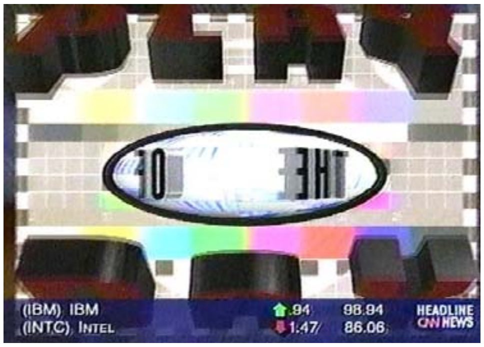


Fig. 3.18 – No rules specifically explain 3D angled text or the reversed text in the center.
`[SPECIAL TEXT]` = `TRUE`; `[DYNAMIC TEXT]` = `TRUE`

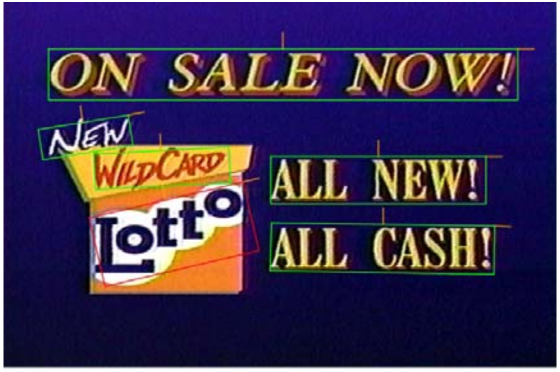


Fig. 3.19 – Non Linear graphic text will be transcribed if `[READABILITY]` = 2. Here

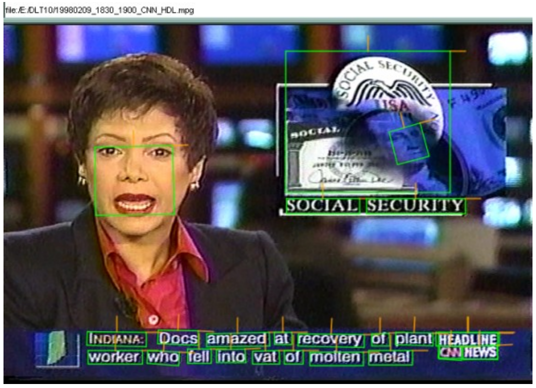

SOCIAL SECURITY
INDIANA: Docs amazed at recovery of plant worker who fell into vat of molten metal
HEADLINE
CNN NEWS

Fig. 3.20 – Non Linear graphic text will not be transcribed if `[READABILITY]` = 0 or 1.

“Lotto” will be transcribed

Here the curved Social Security [LOGO] will not be transcribed

3.3.6 Logos

Text used in Logos is often highly stylized and unorthodox. In such cases, the graphic text must be annotated following the usual text rules but flagged with [LOGO] = TRUE. All other text is [LOGO] =FALSE.

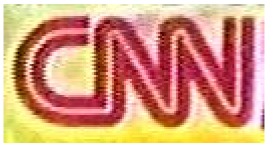


Fig. 3.21 – The CNN [LOGO] contains characters joined in an abnormal fashion. Therefore this text is treated as a graphical text and is annotated with the special attribute [LOGO] = TRUE.

3.4 Hand Annotation Task

3.4.1 Guidelines for Hand Region bounds

The five fingers and the area between the base of the fingers and the wrist define a hand. A hand is considered [VISIBLE] = TRUE if at least 50% of the hand region is present in the scene. A Hand will be annotated using the ‘Point’ data type. When a hand appears a point is placed at the center of the palm region of the hand. The hand is then tracked through successive I-Frames. Examples of Hand annotation are shown in the figures below.

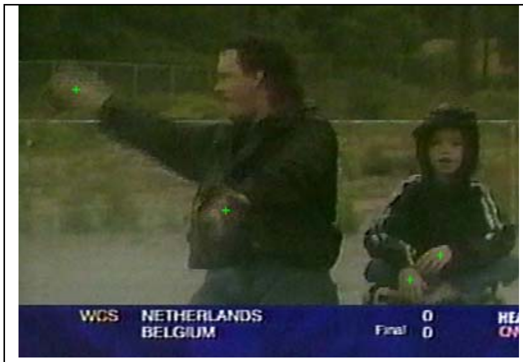
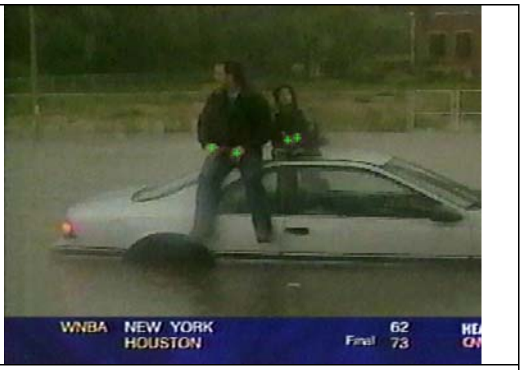
	

Fig. 3.22	Fig. 3.23

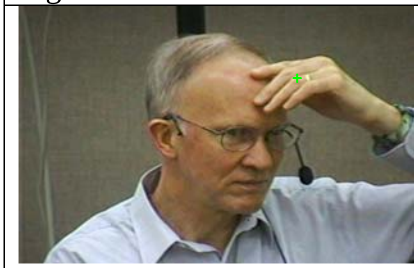
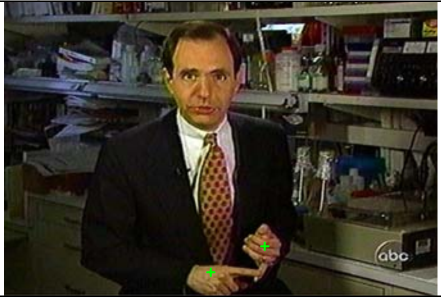

	

Fig. 3.24.1	Fig. 3.24.2
The above figures give example of the hand annotation task in meeting room and broadcast domains.	

A Hand is usually annotated using a 'point'. However, if a hand occupies more than 25% of the frame (more than 160x120 pixels in a 640x480 frame), then the hand is bound with an oriented box denoted as `[LARGE HAND]`. Refer to table in section 2.4.4 for more information.

3.4.2 Hand Task `[VISIBLE]` Attribute

A Hand is considered `[VISIBLE]` when it meets the following three criteria's namely:

- At least 50% of the palm region (includes front or back of hand) can be seen clearly.
- However, if the palm region is seen but the fingers are occluded, `[VISIBLE]` is `TRUE`.
- Conversely, if the fingers are seen without the palm region shown, `[VISIBLE]` is set to `FALSE`.

3.4.3 Other Hand attributes

The `[SYNTHETIC]` attribute of the hand is marked `TRUE` when the hand is not real such as a statue, pictures, etc. Otherwise `[SYNTHETIC]` is set to `FALSE`.

The `[HANDGEAR]` attribute is set to `TRUE` when a hand is covered by something such as gloves. An uncovered/bare hand will have `[HANDGEAR]` set to `FALSE`.

There are a number of scenarios when a Hand is marked with `[OCCLUSION] = TRUE` based on the rules mentioned below.

Rules for Hand Annotation:

- When more than 50% of the hand region is hidden or covered by an object present in the scene the `[OCCLUSION]` attribute of the hand is set to `TRUE` and the `[VISIBLE]` attribute is set to `TRUE`.
- If an object in the scene other than the hand itself hides 50% of the palm region, the Hand will be marked with the `[VISIBLE]` attribute is set to `TRUE` and `[OCCLUSION]` set to `TRUE`.
- When the hands are merged or crossed with each other, the annotator makes a decision based on the features clearly seen, the camera view, etc. This is the annotator's call to decide whether a hand is occluded or not occluded. As seen in the example below in Fig. 3.25.1 the hand at the top is clearly marked with `[VISIBLE]` set to `TRUE` and `[OCCLUSION]` set to `FALSE`. But the hand at the bottom clearly will have `[VISIBLE]` set to `TRUE` and `[OCCLUSION]` set to `TRUE` as the palm region is not seen at all.
- The Hand will be considered with `[VISIBLE]` set to `TRUE` if any one of the six features is blocked by itself as seen in Fig. 3.25.2. The six features include the five fingers and the complete palm region.
- If a Hand was clearly seen and then becomes occluded by something in the scene then the Hand will be marked with `[VISIBLE]` set to `TRUE` and `[OCCLUSION]` set to `TRUE`.
- If a Hand was clearly `[VISIBLE]` and then becomes occluded such that only the fingers are seen in the frame, the Hand is marked with the `[VISIBLE]`set to '`TRUE`' and `[OCCLUSION]` set to '`TRUE`'.

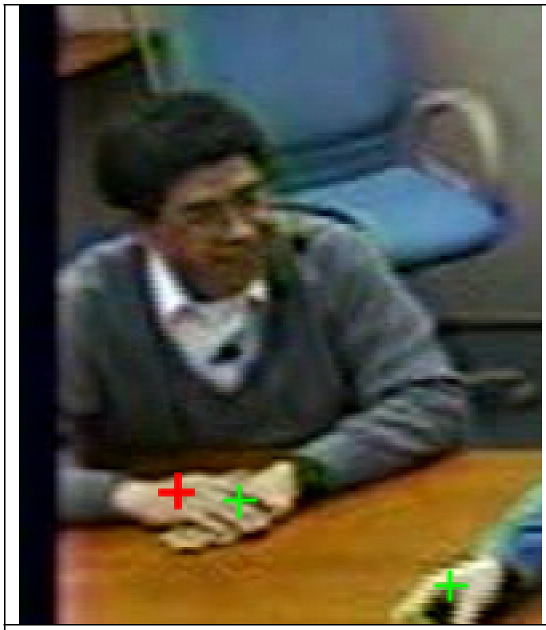
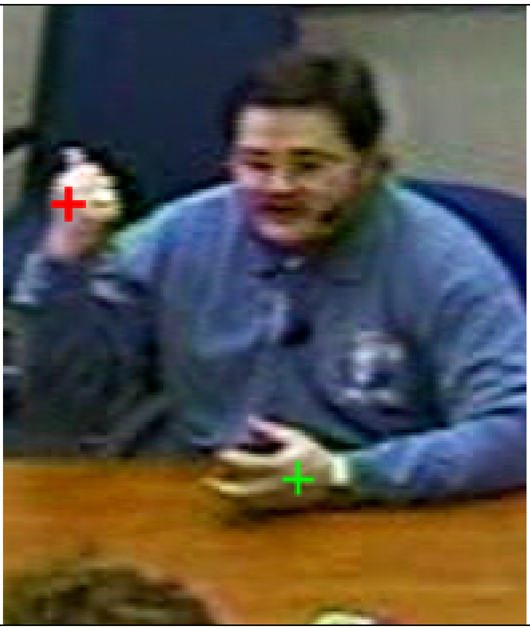
	

Fig. 3.25.1 - The hands marked in green are [OCCLUSION] = TRUE. The hand marked in red is [OCCLUSION] = FALSE.	Fig. 3.25.2 - The hand marked in red occludes itself. Both hands are marked as [OCCLUSION] = FALSE.

3.5 Vehicles Annotation Task

A vehicle is defined as a wheeled or tracked motorized device used to transport occupants. This includes cars, trucks, motorcycles, snowmobiles, tanks, and tractors. In some instances a vehicle will be seen with parts missing or is otherwise incapable of functioning. A vehicle is not annotated until it becomes [VISIBLE]. These will still be annotated so long as the frame of the vehicle is apparent.

A [MOBILITY] attribute has been included to help quantify the motion of the vehicle. A moving vehicle will have [MOBILITY] set to MOBILE while a still vehicle will have [MOBILITY] set to STATIONARY.

A vehicle also contains a [CATEGORY] attribute. This defines the type of the vehicle being annotated. Automobiles and pickup trucks are denoted with [CATEGORY] = 0. Large trucks that contain a covered storage area such as "U-Hauls" and tractor trailers are considered [CATEGORY] = 1. All other vehicles such as motorcycles, golf carts, tanks, bulldozers, will be marked as [CATEGORY] = 2.

3.5.1 Guidelines for Vehicle bounds

Vehicles will be bound using an oriented box. The sides of the box should be in contact with the sides of the vehicle. When a vehicle has become [VISIBLE] and is marked, the box size and orientation will be adjusted to track the vehicle through the scene.

A Vehicle will be annotated only after it has become [VISIBLE] as per the rules in Table 3.1.

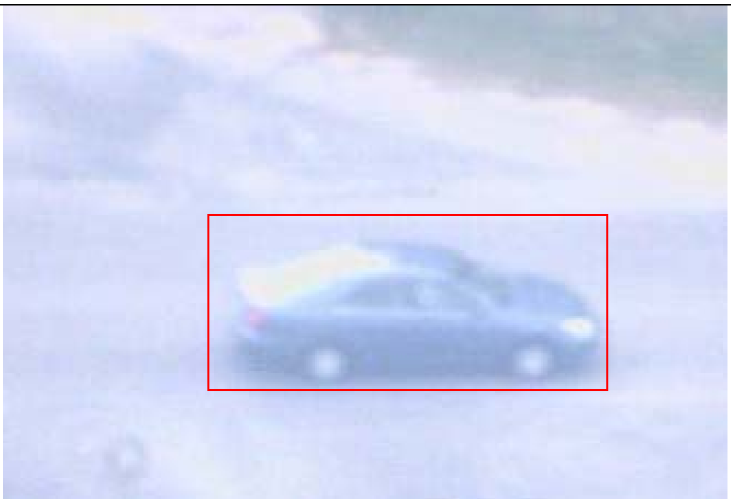


Fig. 3.26 – Sample view of a vehicle from UAV domain

3.5.2 Vehicle Task [VISIBLE] Attribute

1. We change the [ID] of vehicles only when they go out of screen and appear again.
2. To maintain consistency across annotators, we have quantified the camera views as follows and also established clear guidelines for determining visibility:
   a. **Ground Level Camera View**: This includes video footage captured by a camera mounted on the shoulder / hand. Majority of the Broadcast news data contains this camera view.
   b. **Aerial Camera View**: This includes video footage from airplanes, UAV, etc.

Type of Vehicle	[VISIBLE] = TRUE when:	Comments
Hood + Roof + Trunk	1. Hood + Roof completely [VISIBLE] OR 2. Roof + Trunk completely [VISIBLE]	This covers most sedans, cars, passenger trucks
Hood + Roof	1. Hood + Roof/2 OR 2. Complete Roof is [VISIBLE]	This covers SUV and similar vehicles
Roof	1. More than 50% of roof is [VISIBLE]	This covers Vans and Large Trucks

Table 3.1 – Camera View Vehicle [VISIBLE] attribute determination

3.5.3 Description of [GROUP OF VEHICLES] Attribute

Vehicle size varies greatly depending on level of magnification. Often a great number of vehicles can be seen at once, such as in a parking lots as shown in the following figures. In cases where seven or more vehicles are seen and tightly packed together, we mark the group with an oriented box called [GROUP OF VEHICLES]. Bounding for [GROUP OF VEHICLES] will be drawn such that the box will be in contact with the outer edges of the group. Refer to the examples below for instances to use the [GROUP OF VEHICLES] box.

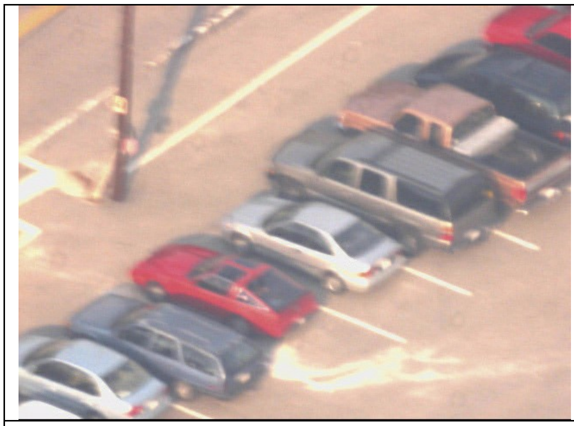
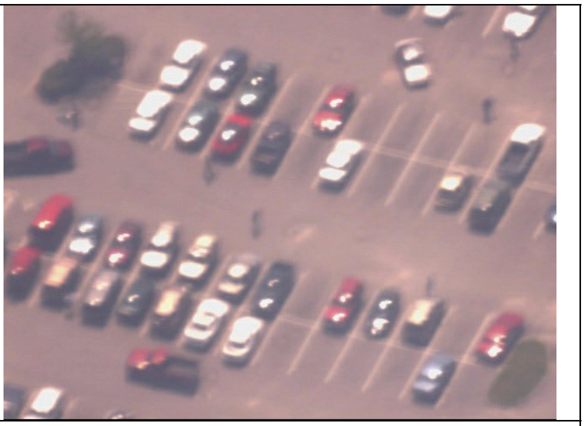
	

Fig. 3.28 – Row of vehicles	Fig. 3.29 – Parking lot full of vehicles

3.5.4 Moving Vehicles in Surveillance Domain

A vehicle is considered to be [PRESENT] when atleast 25% of the vehicle frame is visible. If more than 50% of the vehicle is occluded, the vehicle will have the [OCCLUSION] set to TRUE .A situation when the vehicle frame is cut off by the camera view, the [OCCLUSION] attribute will be set to ‘FALSE’.

The [AMBIGUOUS] attribute has been included to handle extreme occlusion and confusing situations. For e.g. a region where there are many cars which are occluded by the trees and also when the camera view is not clear. A region that is tough to annotate will be tagged with an [AMBIGUOUS] attribute set to TRUE.

Another important difference is the addition of the [MOBILITY] attribute to describe the motion of a vehicle. When a vehicle is moving, [MOBILITY] is set to MOBILE while, if the vehicle is parked or stationary or parked at a particular spot, [MOBILITY] is set to STATIONARY.

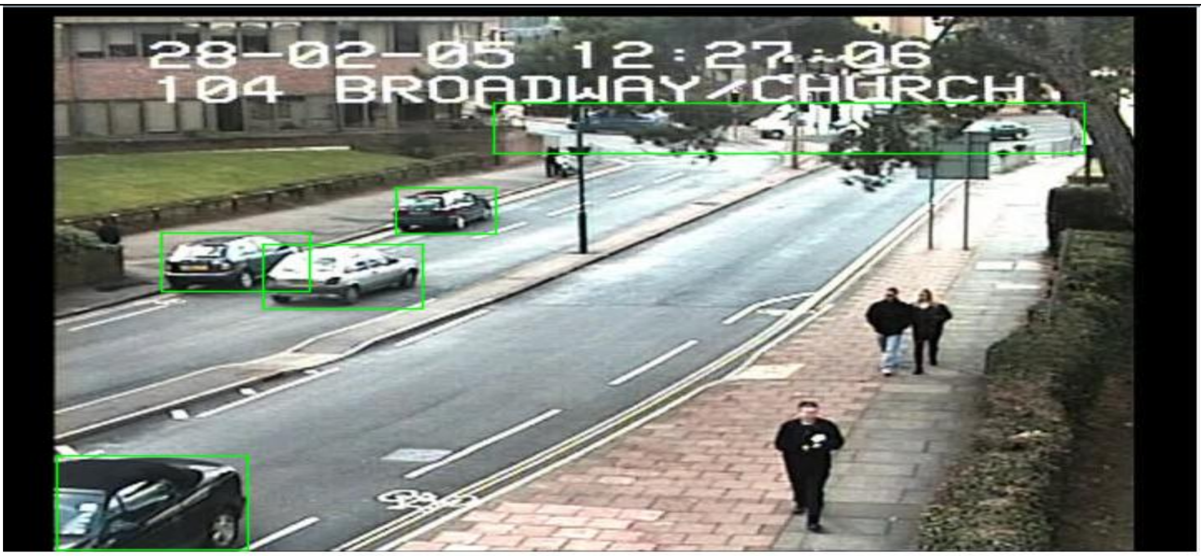


Fig 3.30-Showing the Ambiguous Region with moving vehicles.

3.6 Person Annotation Task

A person is defined as a human entity composed of a head and torso. The current annotation does not include arms and legs as part of the ground truth. An elliptical region will represent the head region of the person and an oriented bounding box will encompass the head region and the upper torso region of the person. The orientation of the bounding region is parallel to the shoulder plane of the person irrespective of the body posture. This again was done to provide consistency and repeatability across multiple annotators.

A person is annotated as soon as any portions of their head or shoulders are seen in the scene and have their [VISIBLE] attribute set to TRUE. In addition to the [VISIBLE] attribute, the person object also includes the [OCCLUSION] attribute which is sent based on the amount of features that are hidden or not seen in the scene.

Finally, an [AMBIGUOUS] attribute has been included to handle extreme occlusion and confusing situations. For e.g. in the meeting room domain, some people are cutoff by the camera view or are almost completely hidden behind a file cabinet. In these cases, [AMBIGUOUS] will be set to TRUE. Table 3.2 gives more detailed information about the means used to determine the different scenarios.

In the UAV domain, a person is defined as a single entity encompassing the whole body. No differentiation is made between the head or torso region as in the meeting room domain.

Another important difference in the UAV domain is the addition of the [MOBILITY] attribute to describe the motion of a person. When a person is moving, [MOBILITY] is set to MOBILE while, if the person is standing or sitting at a particular place, [MOBILITY] is set to STATIONARY.

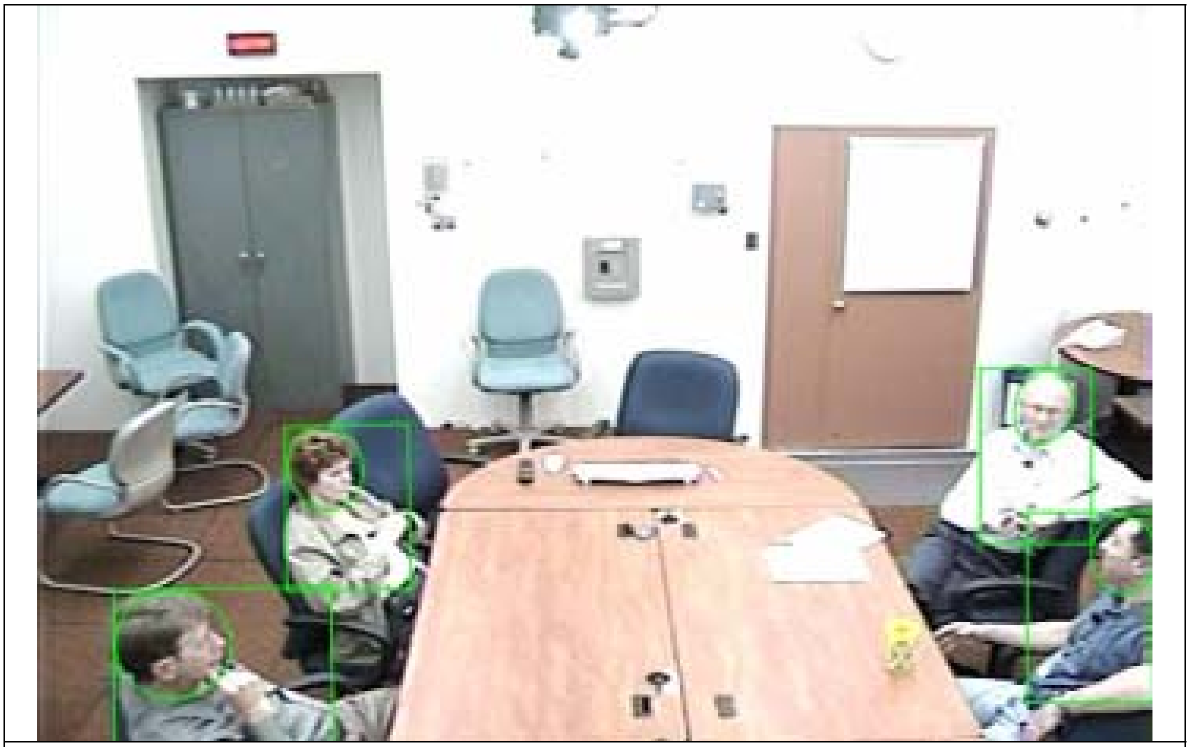


Fig. 3.31 – Snapshot of Person Annotation

3.6.1 Guide for Person Region Bounds

For UAV domain, a person is bound using an oriented bounding box to cover the entire frame of the person.

In Meeting Room domain, a person object is defined by two main regions:

1. An oriented elliptical region covering the complete head.
2. An oriented bounding box covering the head and torso. The bounding box is always drawn vertical to the shoulder plane irrespective of the posture of the person as shown in Fig. 3.31.

3.6.2 Person Task `[VISIBLE]` Attribute

A person will be considered present in the scene if any portion of the Head or Shoulder region can be seen. Hence the `[VISIBLE]` attribute will be set to `TRUE`.

The following table describes additional scenarios as well.

Scenario	`[OCCLUSION]`	`[AMBIGUOUS]`
Both shoulders occluded	-	TRUE
Small portion of head and shoulders occluded	TRUE	-
Major portion of head or shoulders occluded / cutoff	-	TRUE

Table 3.2 – Quantification of `[OCCLUSION]` and `[AMBIGUOUS]` attributes for person in meeting room domain

3.6.2 Person in Surveillance Domain

A person is defined as a human entity composed of a head, torso and legs. The annotation for the surveillance domain includes arms and legs as part of the ground truth. A non-oriented bounding box will be drawn to cover the head, torso and leg region of the person. The non-orientated bounding box will be drawn parallel to the frame. This was done to provide consistency and repeatability across multiple annotators.

A person will be annotated as soon as the annotator sees atleast 25% of the person's body and will have their [PRESENT] attribute set to TRUE. In addition to the [PRESENT] attribute, the person object also includes the [OCCLUSION] attribute, which is based on the amount of features that are hidden or not seen in the scene. If more than 50% of the person body frame is hidden by another object in the frame, the [OCCLUSION] attribute will be set to TRUE.

An `[AMBIGUOUS]` attribute has been included to handle extreme occlusion and confusing situations. When there are a group of people being hidden by an object in the frame or if a person Is not clearly visible, an ambiguous region is defined and any person in this region will be ignored.

Finally, the addition of the `[MOBILITY]` attribute is used to describe the motion of a person in the frame. When a person is moving, the `[MOBILITY]` attribute is set to `MOBILE` while, if the person is standing or sitting at a particular place, `[MOBILITY]` is set to `STATIONARY`.

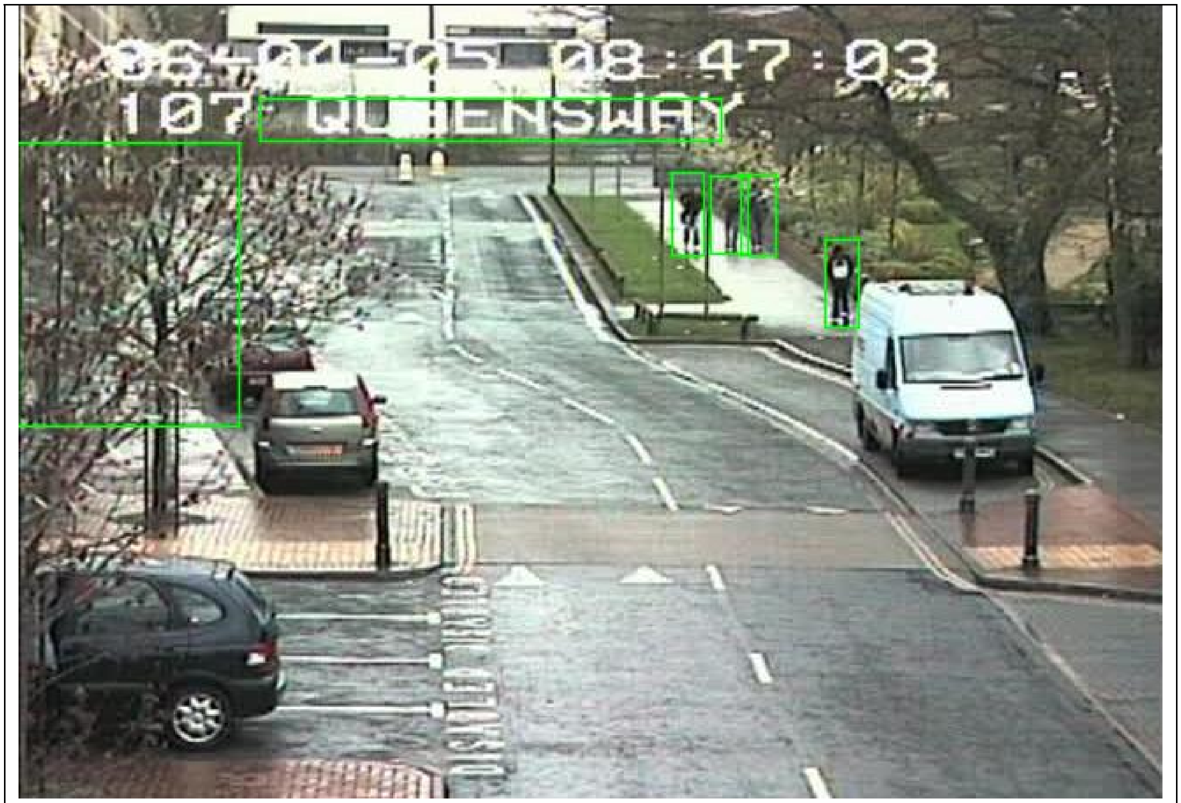


Fig 3.32 -Showing an Ambiguous region with people walking around.

4 Appendix

4.1 ViPER File Format

The ViPER software tool uses a simple XML format for data exchange. It was originally defined for evaluation of text detection, face detection, and person detection.

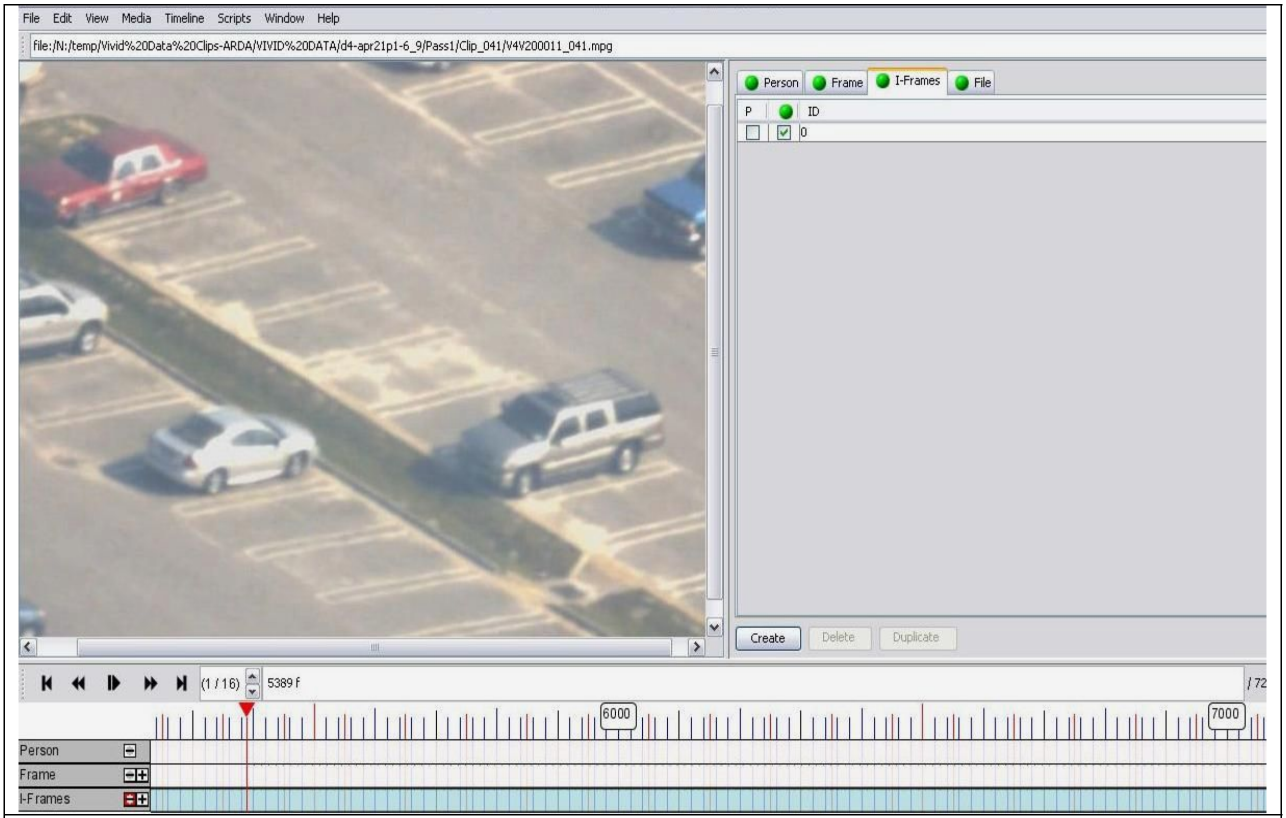
There is a `CONFIG` section which defines the various control descriptors while the `DATA` section instantiates descriptors for one or more media files.

ViPER Hierarchy

1. viper
    1. config
        - descriptor definition
        - descriptor definition
    2. data
        - sourcefile
            - descriptor
        - sourcefile
            - descriptor

Table 4.1 - Overview of the ViPER file hierarchy.

Structural schema: http://viper-toolkit.sourceforge.net/owl/viper/structure
Data schema: http://viper-toolkit.sourceforge.net/owl/viper/datatypes



file:/N:/temp/Vivid%20Data%20Clips-ARDA/VIVID%20DATA/d4-apr21p1-6_9/Pass1/Clip_041/V4V200011_041.mpg

P	●	ID
□	☑	0

Fig. 4.1 – Screenshot of ViPER –GT with a video loaded on the top left pane, time line section at the bottom and the spreadsheet view on the right side.

4.2 Document Naming and Type Conventions

Sl. No.	Type	Description
1.	`[ATTRIBUTE]`	Annotation attributes. E.g. `[VISIBLE]`. Font = Courier New, size 12
2.	`ATTRIBUTE_VALUE`	Annotation attribute or Data type. E.g. `TRUE, MOBILE, BVALUE, DVALUE, CHECKBOX`
3.	General Text	Description of attributes, general text areas. Font = Times New Roman, size 12

4.	X.Y.Z	X = Chapter, Font = Times New Roman, Heading 1
		Y = Section, Font = Times New Roman, Heading 2
		Z = Sub section, Font = Times New Roman, Heading 3

5.	Fig X.Y	Indicates Figure. X corresponds to the chapter number. Y corresponds to the figure number in the given chapter X.
6.	Table X.Y	Indicates Table. X corresponds to the chapter. Y corresponds to the table number in the given chapter X.
7.	`Domain Names`	Domain Names and Annotation Tasks. E.g. `Broadcast News, UAV, Face, Hands, Person` Font: Courier New, Size 12

Table 4.2 Document Naming Conventions

5. References

[1] D. Doermann and D. Mihalcik. Tools and Techniques for video performance evaluation. In *ICPR*, volume 4, pages 167–170, 2000

[2] R. Kasturi, D. Goldgof, P. Soundararajan, and V. Manohar. Performance Evaluation protocol for Text and Face Detection and Tracking in VACE II

[3] R. Kasturi, D. Goldgof, P. Soundararajan, and V. Manohar. Supplement document for the performance evaluation for text and face detection and tracking

[4] H. Raju and S. Prasad from Advanced Interfaces Inc. Video Mining Service Software Suite and documentation

Change Log

4.1 – Updated Person annotation guidelines (Section 3.6)
4.2- Added additional tables for Person and Vehicle in the UAV Domain.
Also, added a brief summary of the new attribute (`[MOBILITY]`) in each section. Updated the ViPER-GT picture in the data schema section.
5.0 – Added new figures, updated guidelines, fixed typos.

6.0 – Major revision of all content, grammatical changes, content heading changes
6.1 -

- Added Tables for Vehicle and Person in Surveillance domain.
- Added description for Vehicle in Surveillance Domain
- Added description for Person in Surveillance Domain.