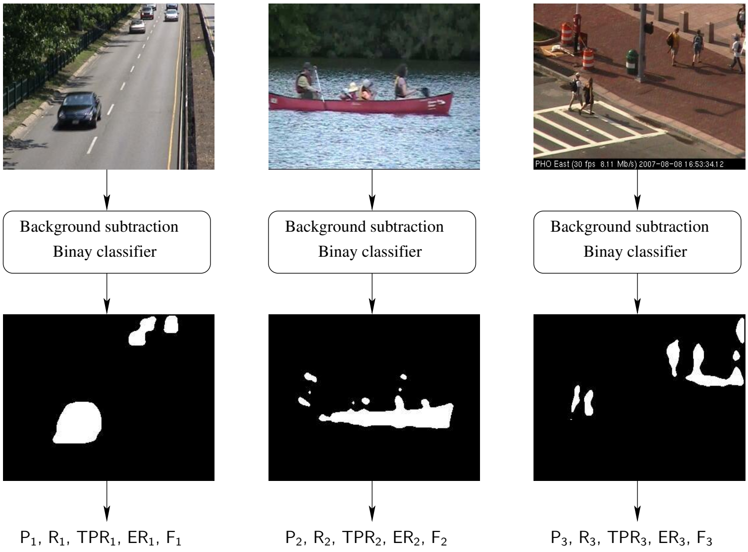
Summarizing the performances of a background subtraction algorithm measured on several videos

Sébastien Piérard and Marc Van Droogenbroeck

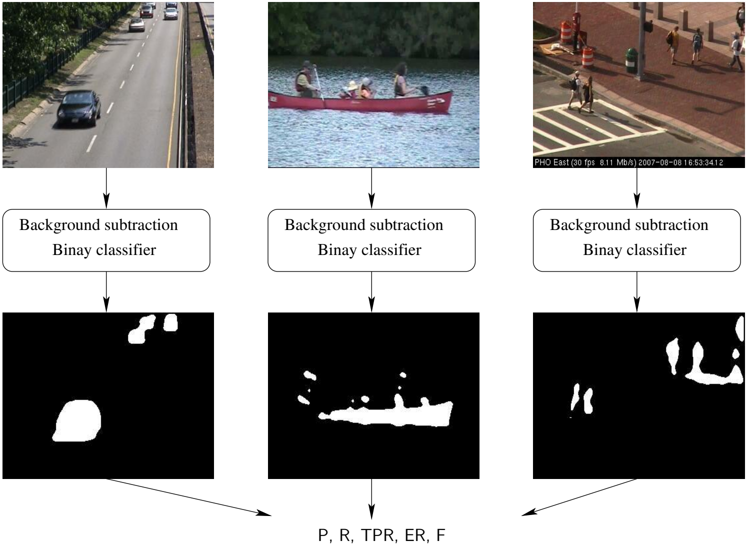
Department of Electrical Engineering and Computer Science (Montefiore Institute), University of Liège, Belgium

Special Session on “Dynamic Background Reconstruction/Subtraction for Challenging Environments”

Motivation: scoring an algorithm for multiple videos



Scoring multiple videos with a unique series of indicators



Should we use the mean for scoring multiple videos?

Do we have a candidate mechanism for aggregating scores of multiple videos?

A natural/obvious candidate for scoring multiple videos is the (arithmetic) mean.

So, if we have:

- Performance for video 1: P_1, R_1, F_1
- Performance for video 2: P_2, R_2, F_2

we calculate

$$\bar{P} = \frac{P_1 + P_2}{2}, \ \bar{R} = \frac{R_1 + R_2}{2}, \text{ and } \bar{F} = \frac{F_1 + F_2}{2}$$

But there is a catch ...

Should we use the mean for scoring multiple videos?

Do we have a candidate mechanism for aggregating scores of multiple videos?

A natural/obvious candidate for scoring multiple videos is the (arithmetic) mean.

So, if we have:

- Performance for video 1: P_1, R_1, F_1
- Performance for video 2: P_2, R_2, F_2

we calculate

$$\bar{P} = \frac{P_1 + P_2}{2},\ \bar{R} = \frac{R_1 + R_2}{2},\ \text{and}\ \bar{F} = \frac{F_1 + F_2}{2}$$

But there is a catch ...

We should not use the mean for scoring multiple videos!

Obviously,

$$\text{for any video } i, \quad \mathsf{F}_i = 2\frac{\mathsf{P}_i \times \mathsf{R}_i}{\mathsf{P}_i + \mathsf{R}_i} \qquad \text{but} \qquad \bar{\mathsf{F}} \neq 2\frac{\bar{\mathsf{P}} \times \bar{\mathsf{R}}}{\bar{\mathsf{P}} + \bar{\mathsf{R}}} = \bar{\bar{\mathsf{F}}}$$

We should not use the mean for scoring multiple videos!

Obviously,

$$\text{for any video } i, \quad \mathsf{F}_i = 2\frac{\mathsf{P}_i \times \mathsf{R}_i}{\mathsf{P}_i + \mathsf{R}_i} \qquad \text{but} \qquad \bar{\mathsf{F}} \neq 2\frac{\bar{\mathsf{P}} \times \bar{\mathsf{R}}}{\bar{\mathsf{P}} + \bar{\mathsf{R}}} = \bar{\bar{\mathsf{F}}}$$

The M4CD algorithm of CDNet 2014 topically illustrates the problem

$$\bar{\mathsf{F}} = 0.69 \qquad \neq \qquad \bar{\bar{\mathsf{F}}} = 2\frac{\bar{\mathsf{P}} \times \bar{\mathsf{R}}}{\bar{\mathsf{P}} + \bar{\mathsf{R}}} = 0.75$$

In fact, the arithmetic mean has severe drawbacks:

- it breaks the intrinsic relationships between probabilistic indicators.
- because of these inconsistencies, we might have that

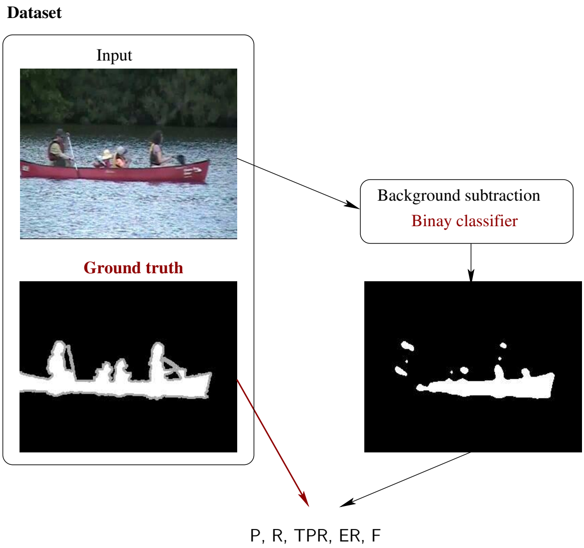
$$\bar{\mathsf{F}}_1 < \bar{\mathsf{F}}_2$$

while

$$\bar{\bar{\mathsf{F}}}_1 > \bar{\bar{\mathsf{F}}}_2 \ .$$

Outline

A scenario for the evaluation of background subtraction algorithms



Towards performance indicators applicable to a binary classifier

Ground truth

Prediction

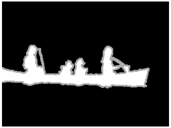
$y \in \{c^+ = \text{foreground}, c^- = \text{background}\}$

$\hat{y} \in \{c^+ = \text{foreground}, c^- = \text{background}\}$

P, R, TPR, ER, F

The confusion matrix

Ground truth

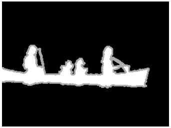


Output

		Predicted class $\hat{y}$	
		Positive	Negative
		TP	FN
		FP	TN

The confusion matrix

Ground truth

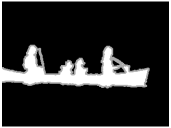


Output

		Predicted class $\hat{y}$	
		Positive	Negative
		TP	FN
		FP	TN

The confusion matrix

Ground truth

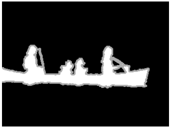


Output

		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	TP	FN
	Negative	FP	TN

The confusion matrix

Ground truth

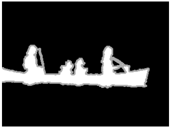


Output

		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	TP	FN
	Negative	FP	TN

The confusion matrix

Ground truth

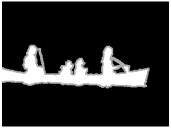


Output

		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	TP	FN
	Negative	FP	TN

The confusion matrix

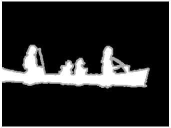
Ground truth



Output

		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	TP	FN
	Negative	FP	TN

Experimental performance indicators based on the confusion matrix I

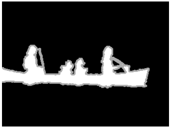
<table>
<tr><td colspan="2"></td><td colspan="2">Predicted class $\hat{y}$</td></tr>
<tr><td colspan="2"></td><td>Positive</td><td>Negative</td></tr>
<tr><td rowspan="2">Actual class y</td><td>Positive</td><td>TP</td><td>FN</td></tr>
<tr><td>Negative</td><td>FP</td><td>TN</td></tr>
</table>

Positive prior $\pi^+ = \frac{\mathrm{TP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}}$

Precision $\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = \mathrm{PPV}$ Positive Predictive Value

True Positive Rate $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \mathrm{R}$ Recall

Experimental performance indicators based on the confusion matrix II



		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	TP	FN
	Negative	FP	TN

Accuracy $$A = \frac{TP+TN}{TP+FN+FP+TN}$$

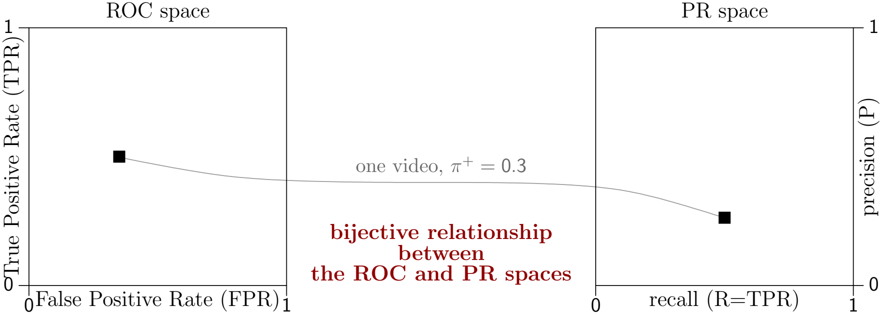
Error rate $$ER = \frac{FP+FN}{TP+FN+FP+TN}$$

F score $$F = \frac{2TP}{2TP+FN+FP}$$

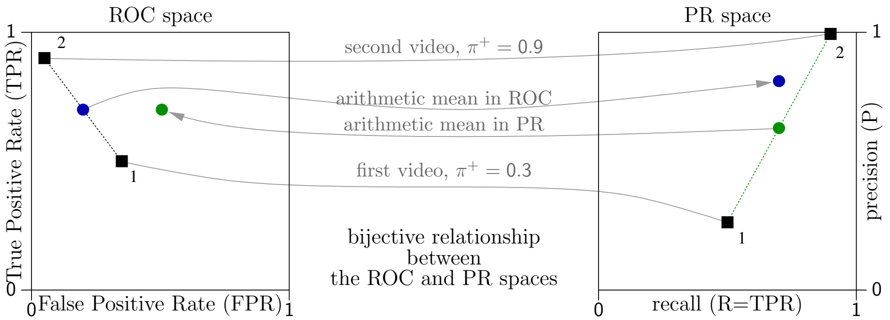
ROC vs PR evaluation spaces: there is a bijection!

There are two well-known evaluation spaces:

ROC: Receiver Operating Characteristic, defined by (FPR, TPR)
PR: Precision/Recall

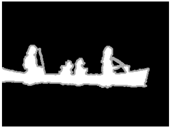


Effect of the arithmetic mean



There is no bijection between the means anymore!

The "normalized" confusion matrix



		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	pTP	pFN
	Negative	pFP	pTN

The proportion of TP, denoted by pTP, is defined as

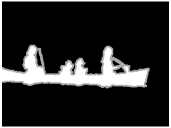
$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}}$$

This has no impact on the calculation of indicators, such as the F score:

$$F = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} = \frac{2\mathrm{pTP}}{2\mathrm{pTP}+\mathrm{pFN}+\mathrm{pFP}}$$

but it leads to a helpful interpretation of experimental indicators in terms of probabilities.

The "normalized" confusion matrix

		Predicted class $\hat{y}$	
		Positive	Negative
Actual class y	Positive	pTP	pFN
	Negative	pFP	pTN

The proportion of TP, denoted by pTP, is defined as

$$\frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

This has no impact on the calculation of indicators, such as the F score:

$$\text{F} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} = \frac{2\text{pTP}}{2\text{pTP} + \text{pFN} + \text{pFP}}$$

but it leads to a helpful interpretation of experimental indicators in terms of probabilities.

Probabilistic meaning of experimental performance indicators

Definition (Joint random experiment for one video)

Draw one pixel at random (all pixels being equally likely) from the video and jointly observe the ground-truth class Y and the predicted class $\hat{Y}$ for this pixel.

Joint random experiment $\Delta = (Y, \hat{Y})$		Prediction $\hat{Y}$	
		Positive	Negative
Ground truth Y	Positive	$\mathrm{tp} = (c^+, c^+)$	$\mathrm{fn} = (c^+, c^-)$
	Negative	$\mathrm{fp} = (c^-, c^+)$	$\mathrm{tn} = (c^-, c^-)$

There are four possible outcomes: $\{\mathrm{tp}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}\}$.

Probabilistic indicators

Joint random experiment $\Delta = (Y, \hat{Y})$		Prediction $\hat{Y}$	
		Positive	Negative
Ground truth Y	Positive	$\mathrm{tp} = (c^+, c^+)$	$\mathrm{fn} = (c^+, c^-)$
	Negative	$\mathrm{fp} = (c^-, c^+)$	$\mathrm{tn} = (c^-, c^-)$

The family of *probabilistic indicators* can be defined based on this random experiment:

$$P\left(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}\right) \text{ with } \emptyset \subsetneq \mathcal{A} \subsetneq \mathcal{B} \subseteq \{\mathrm{tp}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}\} \tag{1}$$

It includes

- $\pi^+ = P\left(\Delta \in \{\mathrm{tp}, \mathrm{fn}\} | \Delta \in \{\mathrm{tp}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}\}\right) = P\left(\Delta \in \{\mathrm{tp}, \mathrm{fn}\}\right)$
- $\mathrm{TPR} = \mathrm{R} = P\left(\Delta = \mathrm{tp} | \Delta \in \{\mathrm{tp}, \mathrm{fn}\}\right)$
- $\mathrm{P} = \mathrm{PPV} = P\left(\Delta = \mathrm{tp} | \Delta \in \{\mathrm{tp}, \mathrm{fp}\}\right)$, $\mathrm{ER} = P\left(\Delta \in \{\mathrm{fn}, \mathrm{fp}\}\right)$
- ... but not the F score!

Probabilistic indicators

<table>
<tr><td colspan="2" rowspan="2">Joint random experiment
$\Delta = (Y, \hat{Y})$</td><td colspan="2">Prediction $\hat{Y}$</td></tr>
<tr><td>Positive</td><td>Negative</td></tr>
<tr><td rowspan="2">Ground truth Y</td><td>Positive</td><td>$\mathrm{tp} = (c^+, c^+)$</td><td>$\mathrm{fn} = (c^+, c^-)$</td></tr>
<tr><td>Negative</td><td>$\mathrm{fp} = (c^-, c^+)$</td><td>$\mathrm{tn} = (c^-, c^-)$</td></tr>
</table>

The family of *probabilistic indicators* can be defined based on this random experiment:

$$P\left(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}\right) \text{ with } \emptyset \subsetneq \mathcal{A} \subsetneq \mathcal{B} \subseteq \{\mathrm{tp}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}\} \tag{1}$$

It includes

- $\pi^+ = P\left(\Delta \in \{\mathrm{tp}, \mathrm{fn}\} | \Delta \in \{\mathrm{tp}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}\}\right) = P\left(\Delta \in \{\mathrm{tp}, \mathrm{fn}\}\right)$
- $\mathrm{TPR} = \mathrm{R} = P\left(\Delta = \mathrm{tp} | \Delta \in \{\mathrm{tp}, \mathrm{fn}\}\right)$
- $\mathrm{P} = \mathrm{PPV} = P\left(\Delta = \mathrm{tp} | \Delta \in \{\mathrm{tp}, \mathrm{fp}\}\right)$, $\mathrm{ER} = P\left(\Delta \in \{\mathrm{fn}, \mathrm{fp}\}\right)$
- ... but not the F score!

Probabilistic indicators

Joint random experiment $\Delta = (Y, \hat{Y})$		Prediction $\hat{Y}$	
		Positive	Negative
Ground truth Y	Positive	$\mathrm{tp} = (c^+, c^+)$	$\mathrm{fn} = (c^+, c^-)$
	Negative	$\mathrm{fp} = (c^-, c^+)$	$\mathrm{tn} = (c^-, c^-)$

The family of *probabilistic indicators* can be defined based on this random experiment:

$$P(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}) \text{ with } \emptyset \subsetneq \mathcal{A} \subsetneq \mathcal{B} \subseteq \{\mathrm{tp}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}\} \tag{1}$$

It includes

- $\pi^+ = P(\Delta \in \{\mathrm{tp}, \mathrm{fn}\} | \Delta \in \{\mathrm{tp}, \mathrm{fn}, \mathrm{fp}, \mathrm{tn}\}) = P(\Delta \in \{\mathrm{tp}, \mathrm{fn}\})$
- $\mathsf{TPR} = \mathsf{R} = P(\Delta = \mathrm{tp} | \Delta \in \{\mathrm{tp}, \mathrm{fn}\})$
- $\mathsf{P} = \mathsf{PPV} = P(\Delta = \mathrm{tp} | \Delta \in \{\mathrm{tp}, \mathrm{fp}\})$, $\mathsf{ER} = P(\Delta \in \{\mathrm{fn}, \mathrm{fp}\})$
- ... but not the F score!

Outline

A probabilistic model for summarization

Definition (Parametric random experiment for several videos)

First, draw one video V at random in the set $\mathbb{V}$, following an arbitrarily chosen distribution $P(V)$. Then, draw one pixel at random from V and observe the ground-truth class Y and the predicted class $\hat{Y}$ for this pixel.

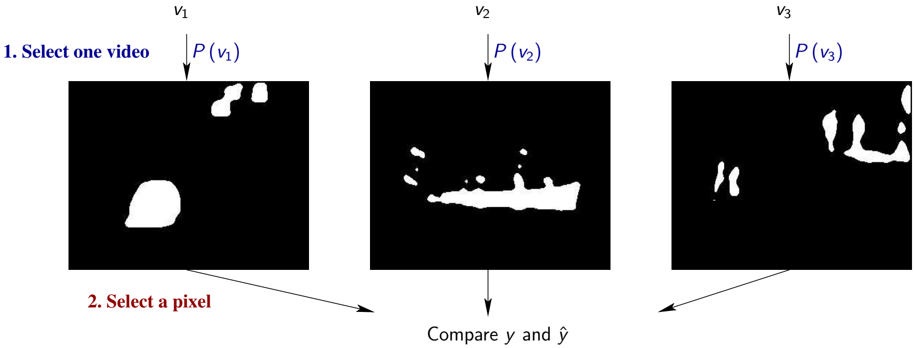


Figure: A probabilistic model for summarization: $\Delta = (V, Y, \hat{Y})$.

Summarization formulas

Notations:

- $\mathrm{I}(v)$ = the value of a performance indicator I for a video $v \in \mathbb{V}$,
- $\mathrm{I}(\mathbb{V})$ = the value of I for a set $\mathbb{V}$ of videos.

We define a probabilistic indicator $\mathrm{I}_{\mathcal{A}|\mathcal{B}}$ as $P\left(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}\right)$, and $\mathrm{I}_{\mathcal{B}}$ as $P\left(\Delta \in \mathcal{B}\right)$. We have

$$\begin{aligned}
\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) &= P\left(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}\right) \\
&= \sum_{v \in \mathbb{V}} P\left(\Delta \in \mathcal{A}, V = v | \Delta \in \mathcal{B}\right) \\
&= \sum_{v \in \mathbb{V}} P\left(V = v | \Delta \in \mathcal{B}\right) P\left(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}, V = v\right) \\
\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) &= \sum_{v \in \mathbb{V}} P\left(V = v | \Delta \in \mathcal{B}\right) \mathrm{I}_{\mathcal{A}|\mathcal{B}}(v)
\end{aligned} \tag{2}$$

For the particular case of an unconditional probabilistic indicator $\mathrm{I}_{\mathcal{A}} = \mathrm{I}_{\mathcal{A}|\{\mathrm{tn},\mathrm{fp},\mathrm{fn},\mathrm{tp}\}}$, we have

$$\mathrm{I}_{\mathcal{A}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} P\left(V = v\right) \mathrm{I}_{\mathcal{A}}(v) \tag{3}$$

Summarization formulas

Notations:

- $\mathsf{I}(v)$ = the value of a performance indicator I for a video $v \in \mathbb{V}$,
- $\mathsf{I}(\mathbb{V})$ = the value of I for a set $\mathbb{V}$ of videos.

We define a probabilistic indicator $\mathsf{I}_{\mathcal{A}|\mathcal{B}}$ as $P(\Delta \in \mathcal{A} | \Delta \in \mathcal{B})$, and $\mathsf{I}_{\mathcal{B}}$ as $P(\Delta \in \mathcal{B})$. We have

$$\begin{aligned}
\mathsf{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) &= P(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}) \\
&= \sum_{v \in \mathbb{V}} P(\Delta \in \mathcal{A}, V = v | \Delta \in \mathcal{B}) \\
&= \sum_{v \in \mathbb{V}} P(V = v | \Delta \in \mathcal{B}) \, P(\Delta \in \mathcal{A} | \Delta \in \mathcal{B}, V = v) \\
\mathsf{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) &= \sum_{v \in \mathbb{V}} P(V = v | \Delta \in \mathcal{B}) \, \mathsf{I}_{\mathcal{A}|\mathcal{B}}(v) \qquad (2)
\end{aligned}$$

For the particular case of an unconditional probabilistic indicator $\mathsf{I}_{\mathcal{A}} = \mathsf{I}_{\mathcal{A}|\{\mathrm{tn},\mathrm{fp},\mathrm{fn},\mathrm{tp}\}}$, we have

$$\mathsf{I}_{\mathcal{A}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} P(V = v) \, \mathsf{I}_{\mathcal{A}}(v) \qquad (3)$$

Summarization formulas and properties

Formulas:

$$\mathrm{I}_{\mathcal{A}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} P(V = v)\, \mathrm{I}_{\mathcal{A}}(v)$$

$$\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} P(V = v | \Delta \in \mathcal{B})\, \mathrm{I}_{\mathcal{A}|\mathcal{B}}(v)$$

$$\text{Example: } \mathrm{TPR}(\mathbb{V}) = \frac{1}{\pi^+(\mathbb{V})} \sum_{v \in \mathbb{V}} P(V = v)\, \pi^+(v)\, \mathrm{TPR}(v) \tag{4}$$

Properties:

1. Summarization preserves the consistency between indicators, including the bijection between the ROC and PR spaces!
2. As long as an indicator is defined for at least one video, it can be summarized! To prove it, we rewrite $\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V})$ as

$$\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) = \frac{\mathrm{I}_{\mathcal{A}\cap\mathcal{B}}(\mathbb{V})}{\mathrm{I}_{\mathcal{B}}(\mathbb{V})} = \frac{\mathrm{I}_{\mathcal{A}\cap\mathcal{B}}(\mathbb{V})}{\sum_{v \in \mathbb{V}} P(V = v)\, \mathrm{I}_{\mathcal{B}}(v)} \tag{5}$$

Summarization formulas and properties

Formulas:

$$\mathrm{I}_{\mathcal{A}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} P(V = v)\, \mathrm{I}_{\mathcal{A}}(v)$$

$$\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) = \sum_{v \in \mathbb{V}} P(V = v | \Delta \in \mathcal{B})\, \mathrm{I}_{\mathcal{A}|\mathcal{B}}(v)$$

$$\text{Example: } \mathrm{TPR}(\mathbb{V}) = \frac{1}{\pi^{+}(\mathbb{V})} \sum_{v \in \mathbb{V}} P(V = v)\, \pi^{+}(v)\, \mathrm{TPR}(v) \tag{4}$$

Properties:

1. Summarization preserves the consistency between indicators, including the bijection between the ROC and PR spaces!
2. As long as an indicator is defined for at least one video, it can be summarized! To prove it, we rewrite $\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V})$ as

$$\mathrm{I}_{\mathcal{A}|\mathcal{B}}(\mathbb{V}) = \frac{\mathrm{I}_{\mathcal{A}\cap\mathcal{B}}(\mathbb{V})}{\mathrm{I}_{\mathcal{B}}(\mathbb{V})} = \frac{\mathrm{I}_{\mathcal{A}\cap\mathcal{B}}(\mathbb{V})}{\sum_{v \in \mathbb{V}} P(V = v)\, \mathrm{I}_{\mathcal{B}}(v)} \tag{5}$$

An algorithm for the computation of summarized indicators

Algorithm:

1. Blend the **normalized** confusion matrices with the $P(v_1)$, $P(v_2)$, ... weights,
2. then calculate the indicators!

$$\begin{array}{cc} \mathrm{pTP}_1 & \mathrm{pFN}_1 \\ \mathrm{pFP}_1 & \mathrm{pTN}_1 \end{array} \qquad \begin{array}{cc} \mathrm{pTP}_2 & \mathrm{pFN}_2 \\ \mathrm{pFP}_2 & \mathrm{pTN}_2 \end{array} \qquad \begin{array}{cc} \mathrm{pTP}_3 & \mathrm{pFN}_3 \\ \mathrm{pFP}_3 & \mathrm{pTN}_3 \end{array}$$

$$P(v_1) \qquad\qquad P(v_2) \qquad\qquad P(v_3)$$

$$\begin{array}{cc} \sum_v P(v)\,\mathrm{pTP}_v & \sum_v P(v)\,\mathrm{pFN}_v \\ \sum_v P(v)\,\mathrm{pFP}_v & \sum_v P(v)\,\mathrm{pTN}_v \end{array}$$

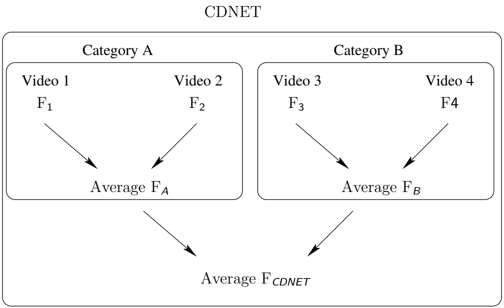
$$\mathrm{F} = \frac{2\mathrm{pTP}}{2\mathrm{pTP} + \mathrm{pFN} + \mathrm{pFP}}$$

Outline

Experiments with CDNET 2014

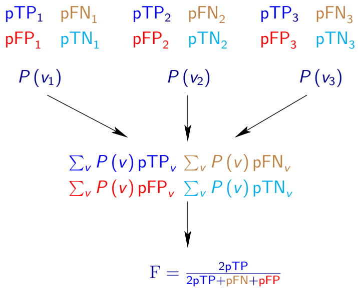
We analyze two scenarios:

The original CDNET procedure



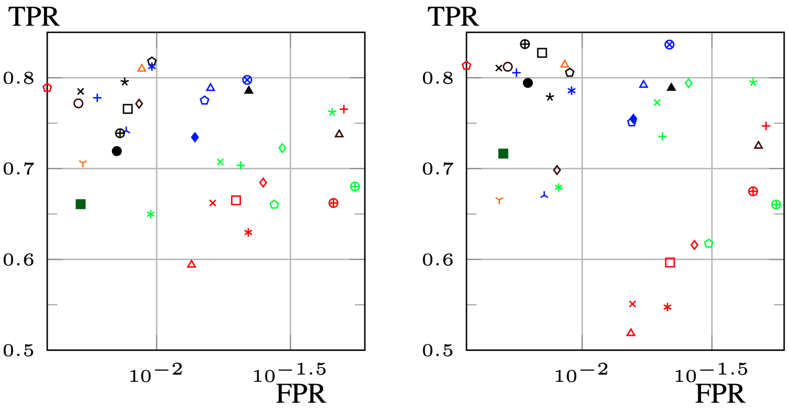
Our summarization, with

$$P(V = v) = \frac{1}{11} \times \frac{1}{M}$$



In the ROC space

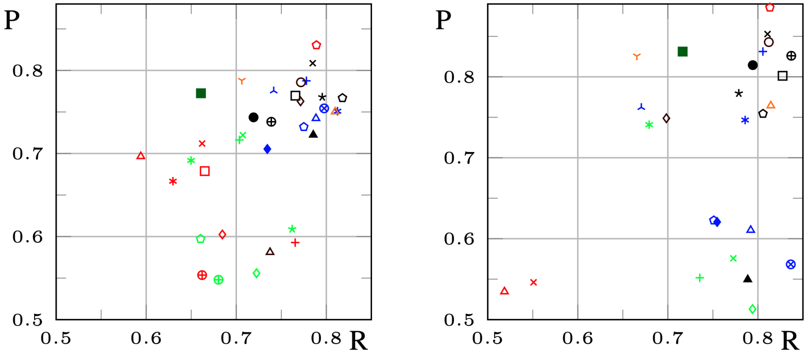
36 classifiers evaluated on the CDNET 2014 dataset in the ROC space:



(a) CDNET procedure (ROC space).   (b) Our procedure (ROC space).

Figure: Summarized performances according to two different procedures in the cropped ROC space.

In the PR space



(c) CDNET procedure (PR space). (d) Our procedure (PR space).

Figure: Summarized performances according to two different procedures in the cropped PR space.

Remember that our summarization procedure preserves the bijection between the ROC and PR evaluation spaces!

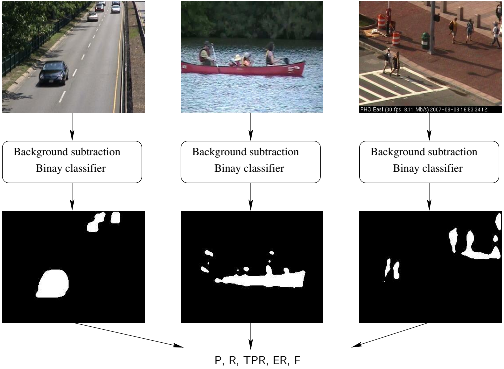
Ranking based on the F scores

Algorithm	F of CDNET 2014	F [our summarization]
SemanticBGS	**0.8098 (1)**	**0.8479 (1)**
IUTIS-5	0.7821 (2)	0.8312 (3)
IUTIS-3	0.7694 (3)	0.8182 (5)
WisenetMD	0.7559 (4)	0.7791 (10)
SharedModel	0.7569 (5)	0.7885 (8)
WeSamBE	0.7491 (6)	0.7792 (9)
SuBSENSE	0.7453 (7)	0.7657 (12)
PAWCS	0.7478 (8)	0.8272 (4)

Table: Extract of F scores (and ranks) obtained with two procedures on CDNET.

Outline

Take-home messages



1. It is unsound to average performance indicators, such as P, TPR, with the arithmetic mean because
   1. it breaks the consistency between indicators
   2. it makes the interpretation less reliable
2. Prefer the summarization formulas

More on summarization:
`http://www.telecom.ulg.ac.be/summarization`